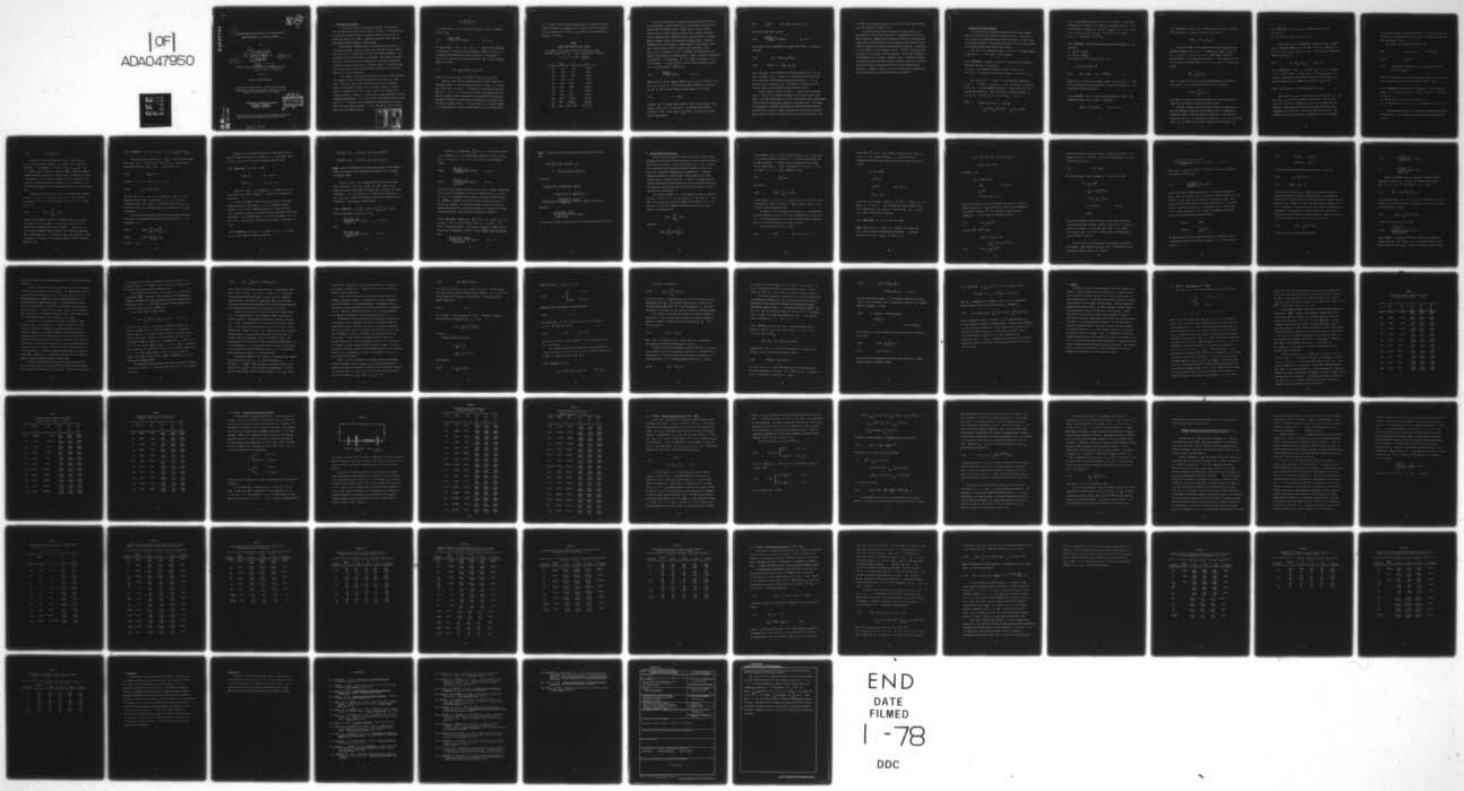
AD-A047 950 STANFORD UNIV CALIF DEPT OF OPERATIONS RESEARCH F/G 12/1
VARIANCE REDUCTION TECHNIQUES FOR THE SIMULATION OF MARKOV PROC--ETC(U)
OCT 77 P HEIDELBERGER N00014-76-C-0578
UNCLASSIFIED TR-42 NL

OF
ADA047950



END
DATE FILMED
1-78
DDC

VARIANCE REDUCTION TECHNIQUES FOR THE SIMULATION OF MARKOV PROCESSES. I. MULTIPLE ESTIMATES.

by

Philip Heidelberger*

TECHNICAL REPORT NO. 42

October 1977

76p.

TR-42

Prepared under Contract N00014-76-C-0578 (NR 042-343)

NSF-MCS-75-23607

for the

Office of Naval Research

DEPARTMENT OF OPERATIONS RESEARCH
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

*The research of this author was also partially supported under National Science Foundation Grant MCS75-23607.

1. Introduction and Summary

In recent years computer simulation has become a very important tool for analyzing the behavior of stochastic processes. As the structures of widely used processes become increasingly complex, analytic results become more difficult to obtain. Frequently simulation is the only computationally feasible method to study a process.

Unfortunately, simulation can be a very expensive tool to use. It is therefore desirable to develop methods that can reduce the run lengths (and hence cost) of a simulation, yet still give accurate estimates. Such methods are called variance reduction techniques. This paper will propose and test a new variance reduction technique in the special case when the stochastic process being simulated is a Markov process. A subsequent paper will describe several other related techniques applicable when the Markov process has a finite state space.

As an example of how expensive simulations can be, consider estimating via simulation $E(W)$, the expected stationary waiting time in an $M/M/1$ queue. The $M/M/1$ queue is not something that one would ordinarily simulate since analytic results for it are readily available. However, despite its simplicity the $M/M/1$ queue can be a very difficult and expensive process to simulate. It is therefore a good candidate for testing simulation methodologies. Let λ be the rate of the Poisson arrival process and let μ^{-1} be the mean of the exponentially distributed service times. Define the usual traffic intensity $\rho = \lambda/\mu$. Now let W_n be the waiting time of the nth customer and let

$$\bar{W}_N = \frac{1}{N+1} \sum_{n=0}^{N} W_n \quad .$$

It is known that if $\rho < 1$ (see Crane and Iglehart (1974a) or Iglehart (1971)), then

$$(1.1) \qquad \frac{\sqrt{N}(\bar{W}_N - E(W))}{\sigma} \Rightarrow N(0,1) \qquad \text{as } N \to \infty$$

for some constant σ $(0 < \sigma < \infty)$. Here $\Rightarrow$ denotes weak convergence or convergence in distribution (see Billingsley (1968)) and $N(0,1)$ is a normally distributed random variable with mean 0 and variance 1. From (1.1) we can form the following approximate $100 \times (1-\alpha)\%$ confidence interval for $E(W)$

$$(1.2) \qquad (\bar{W}_N - z_{\alpha/2}\, \sigma/\sqrt{N},\ \bar{W}_N + z_{\alpha/2}\, \sigma/\sqrt{N})$$

where, for a given α, $z_{\alpha/2}$ is defined by $P\{N(0,1) > z_{\alpha/2}\} = \alpha/2$.

Suppose we want to run our simulation until the half length of our confidence interval is some prespecified fraction of $E(W)$; i.e., we want $z_{\alpha/2}\, \sigma/\sqrt{N} = \delta E(W)$ for some δ. For example if we (somewhat arbitrarily) pick $\alpha = .1$ and $\delta = .1$ we then seek to form a 90% confidence interval whose half length is 10% of the quantity we are interested in estimating. The number of customers, N, that need to be simulated to obtain this level of accuracy is given in Table 1. It is seen that as ρ increases (beyond $\rho = .3$) the required run lengths increase rapidly until for large values

of ρ one must simulate such an enormous number of customers to obtain "decent" estimates that simulation is no longer a feasible alternative. It is run lengths such as these that variance reduction techniques are designed to cut down.

TABLE 1

SAMPLE SIZES FOR THE M/M/1 QUEUE

N = Number of customers that must be simulated for a 90% confidence interval for E(W) to have a half length of .10 E(W)

$$(\mu = 1.0,\ \lambda = \rho,\ E(W) = \frac{\lambda}{\mu(\mu-\lambda)})$$

ρ	E(W)	σ^2	N
.10	.111	.375	8,200
.20	.250	1.39	6,020
.30	.429	3.96	5,830
.40	.667	10.6	6,430
.50	1.00	29.0	7,850
.60	1.50	88.5	10,600
.70	2.33	335	16,700
.80	4.00	1,976	33,400
.90	9.00	35,901	119,000
.95	19.0	607,600	455,000
.99	99.0	3.95×10^8	1.09×10^7

One of the most effective variance reduction techniques is that of control variables. A good introduction to this technique is given in the book by Gaver and Thompson (1973). Recent studies involving control variables may be found in Gaver and Shedler (1971), Iglehart and Lewis (1976), Lavenberg (1974), Lavenberg, Moeller and Sauer (1977), Lavenberg, Moeller and Welch (1977), and Lavenberg and Shedler (1975). Since the technique about to be proposed is closely related to this method, we now give a brief outline of control variables before proceeding. Let $\{X_n, n \geq 0\}$ be a sequence of i.i.d. (independent and identically distributed) random variables with unknown mean $r = E(X_n)$. We shall be interested in estimating r via simulation. Let $\sigma_x^2 = \sigma^2(X_n)$, the variance of X_n. We can estimate r by $\bar{X}_N = \sum_{n=1}^{N} X_n/N$ and can form a confidence interval for r using the central limit theorem

$$\frac{\sqrt{N}(\bar{X}_N - r)}{\sigma_x} \Rightarrow N(0,1) \qquad \text{as } N \to \infty . \tag{1.3}$$

Suppose that we now have random variables $\{C_n, n \geq 0\}$ such that the C_n's are i.i.d., X_n and C_n are correlated (usually achieved by simulating X_n and C_n with the same stream of random numbers) and for which

$$r_c = E(C_n) \tag{1.4}$$

is known. Let β be some constant and set $Z_n(\beta) = X_n + \beta(C_n - r_c)$. Then $\{Z_n(\beta), n \geq 0\}$ are i.i.d. with mean r and variance which will be denoted by $\sigma^2(\beta)$. We let $\bar{Z}_N(\beta) = \sum_{n=1}^{N} Z_n(\beta)/N$. We have by the strong law of large numbers

(1.5) $\bar{Z}_N(\beta) \to r$ a.s. (almost surely) as $N \to \infty$

and by the central limit theorem;

(1.6) $$\frac{\sqrt{N}(\bar{Z}_N(\beta) - r)}{\sigma(\beta)} \Rightarrow N(0,1) \qquad \text{as } N \to \infty .$$

We now pick $\beta = \beta^*$ to minimize the variance term $\sigma^2(\beta)$. It is easy to show that

(1.7) $$\beta^* = -\text{cov}(X_n, C_n)/\sigma^2(C_n) ,$$

(1.8) $$\sigma^2(\beta^*) = (1 - \rho^2(X_n, C_n))\, \sigma_x^2$$

where $\rho(X_n, C_n)$ is the coefficient of correlation between X_n and C_n. Since $0 \le \rho^2(X_n, C_n) \le 1$ a reduction in variance has been obtained and we are thus able to form shorter confidence intervals for r. C_n is called a control variable for X_n. The method can be extended to allow multiple controls (see Lavenberg, Moeller and Sauer (1977)).

The key things to observe are that $r_c = E(C_n)$ must be known and that X_n and C_n must be highly correlated to get large variance reductions. It is often very difficult to come up with good controls, particularly if the stochastic process being simulated is very complicated. The method to be proposed in this paper circumvents this difficulty by devising controls which will usually be highly correlated with the process of interest and for which the means of the controls need not be explicitly known. This

is because the controls are chosen in such a way that their means actually equal the parameter of interest (r).

We shall initially restrict ourselves to studying controls for functionals of the stationary distribution of regenerative discrete time Markov processes. Examples of such processes are positive recurrent Markov chains and the server workload process in multiple server queues in light traffic (from which the waiting time process can be derived). In Section 2 we state results for such processes. Section 3 contains a description and discussion of the variance reduction technique. We also show in Section 3 how the method can be extended to different types of stochastic processes including continuous time Markov chains, semi-Markov processes and certain types of stationary processes (which may be neither regenerative nor Markovian). Numerical examples, taken from queueing theory, demonstrating the effectiveness of the method are presented in Section 4.

2. Regenerative Markov Processes

In this section we introduce notation and state some preliminary results for discrete time regenerative Markov processes. These results will be useful in developing variance reduction techniques for this class of stochastic processes. The notation and definitions in the early portion of this section follow Orey (1971) fairly closely.

Let $(E,\mathcal{E})$ be a measurable space; i.e., let $\mathcal{E}$ be a sigma algebra of subsets of $\mathcal{E}$. Let $\mathbb{R}$ denote the real numbers.

(2.1) DEFINITION. A function $P:(E,\mathcal{E}) \to \mathbb{R}$ is said to be a probability transition function if it satisfies

(a) $P(x,\cdot)$ is a probability measure on $(E,\mathcal{E})$ for all $x \in E$.

(b) $P(\cdot,B)$ is a measurable function with respect to $\mathcal{E}$ for all $B \in \mathcal{E}$.

Let $E^\infty = E \times E \times \cdots$ and let $\mathcal{E}^\infty$ be the product sigma field $\mathcal{E} \times \mathcal{E} \times \cdots$. For any $\omega = (\omega_0, \omega_1, \ldots) \in E^\infty$ let $X_n(\omega) = \omega_n$ for each $n \geq 0$. Let μ be some probability measure on $(E,\mathcal{E})$, called the initial probability distribution. There then exists a probability measure P_μ on $(E^\infty, \mathcal{E}^\infty)$ such that for all $n \geq 0$ and $B_0, \ldots, B_n \in \mathcal{E}$

$$(2.2) \qquad P_\mu\{X_0 \in B_0, X_1 \in B_1, \ldots, X_n \in B_n\}$$

$$= \int_{B_0} \mu(dx_0) \int_{B_1} P(x_0, dx_1) \cdots \int_{B_n} P(x_{n-1}, dx_n) \quad .$$

If Y is any random variable let $E_\mu(Y) = \int Y(\omega)\, P_\mu(d\omega)$ (if the region of integration is omitted it is assumed to be the whole space E). If Y is an indicator function; i.e., if $Y(\omega) = 1_{\{B\}}(\omega)$ ($= 1$ if $\omega \in B$ and 0 if $\omega \notin B$) let $P_\mu(B) = E_\mu(1_{\{B\}})$. If for some $x \in E$ we have $\mu\{\{x\}\} = 1$ we then write E_x for E_μ.

(2.3) DEFINITION. The n-step probability transition functions P^n are defined by

(a) $P^0(x,B) = 1_{\{B\}}(x)$,

(b) $P^1(x,B) = P(x,B)$,

(c) $P^n(x,B) = \int P^{n-1}(x, dy)\, P(y,B)$, for $n \geq 2$.

It can then be shown that

$$(2.4) \qquad P_\mu\{X_n \in B | X_0, \ldots, X_k\} = P^{n-k}(X_k, B) \qquad \text{a.s.}$$

Equation (2.4) is known as the Markov property and $\{X_n, n \geq 0\}$ is said to be a Markov process with state space E, initial distribution μ, and stationary transition function P.

(2.5) DEFINITION. Let π be any probability measure on $(E, \mathcal{E})$. The probability measure πP on $(E, \mathcal{E})$ is defined by

$$\pi P(B) = \int \pi(dx)\, P(x,B), \qquad \text{for all } B \in E .$$

(2.6) DEFINITION. For any $n \geq 0$ and any function $f:E \to \mathbb{R}$ such that $\int P^n(x, dy)\ |f(y)| < \infty$ define the function $P^n f:E \to \mathbb{R}$ by

$$P^n f(x) = \int P^n(x, dy)\ f(y) .$$

If the state space E is countable we can label the states by the nonnegative integers $\{0, 1, 2, \ldots\}$. Letting $P_{ij} = P\{X_{n+1} = j | X_n = i\}$ and $\mu_i = P\{X_0 = i\}$ we have $\{X_n, n \geq 0\}$ being a Markov chain with transition matrix P and initial distribution μ. The integrals in (2.2) to (2.6) can be replaced by sums; e.g., if $x = i$ and $B = \{k\}$ then (2.3.c) becomes

$$P^{n+1}_{ik} = \sum_{j=0}^{\infty} P^n_{ij}\ P_{jk}$$

which is a special case of the familiar Chapman-Kolmogorov equations for Markov chains. In the countable case Definition (2.6) becomes

$$P^n\ f(i) = \sum_{j=0}^{\infty} P^n_{ij}\ f(j) .$$

The reader is referred to the books by Chung (1967) and Karlin and Taylor (1975) for a more detailed analysis of Markov chains.

In order to develop results which will be useful in simulation we make the additional hypothesis that the Markov process is regenerative (see Çinlar (1975) or Crane and Iglehart (1975)); that is we assume that there exists some $x_0 \in E$ such that with probability 1, $X_n = x_0$ infinitely often. Let T_m denote the mth time the process enters the state x_0.

(2.7) DEFINITION. For $m \geq 0$, T_m is defined (recursively) by

(a) $T_0 = 0$,

(b) $T_m = \inf\{n > T_{m-1} : X_n = x_0\}$ for $m \geq 1$.

Notice that T_m is a stopping time (see Çinlar (1975)). Because $X_n = x_0$ infinitely often, $T_m < \infty$ for each m and T_m increases to ∞ as $m \to \infty$. Let $\tau_m = T_m - T_{m-1}$ for $m \geq 1$. For each $m \geq 1$ denote the random vector $\underset{\sim}{V}_m$ by

$$\underset{\sim}{V}_m = (X_{T_{m-1}}, X_{T_{m-1}+1}, \ldots, X_{T_m-1}, \tau_m) . \tag{2.8}$$

(2.9) PROPOSITION. If $X_0 = x_0$ a.s. then $\{\underset{\sim}{V}_m, m \geq 1\}$ is a sequence of i.i.d. random vectors. If $P\{X_0 = x_0\} < 1$ then $\{\underset{\sim}{V}_m, m \geq 1\}$ is a sequence of independent random vectors although the distribution of $\underset{\sim}{V}_1$ may be different than the (common) distribution of $\{\underset{\sim}{V}_m, m \geq 2\}$.

PROOF. See Proposition 1 of Crane and Iglehart (1974b). □

The process is said to be in the mth cycle between times T_{m-1} and $T_m - 1$, and τ_m is called the length of the mth cycle. The fact that $\{\underset{\sim}{V}_m, m \geq 1\}$ are i.i.d. (assuming $X_0 = x_0$ a.s.) just means that the behavior of the process over a cycle has the same distribution and is independent of the behavior of the process over any other cycle. The importance of this in the context of simulation is that the simulation

run can then be broken up into randomly spaced i.i.d. blocks so that the techniques of classical statistics can be used to analyze the output (see Crane and Iglehart (1975)).

We now make the additional assumptions that

(2.9) $$P_y\{\tau_1 < \infty\} = 1 , \qquad \text{for all } y \in E$$

(2.10) $$E_{x_0}(\tau_1) < \infty$$

(2.11) The (common) distribution of $\{\tau_m, m \geq 2\}$ is aperiodic (see Çinlar (1975)).

Under these conditions the following proposition (whose proof may be found in either Çinlar (1975) or Crane and Iglehart (1975)) is true.

(2.12) PROPOSITION. There exists a random element X and a probability measure π on $(E, \mathcal{E})$ with $P\{X \in A\} = \pi(A)$ for all $A \in \mathcal{E}$ such that

(a) $X_n \Rightarrow X$,

(b) $P^n(y,A) \to \pi(A)$ for all $A \in \mathcal{E}$ and $y \in E$,

(c) π satisfies $\pi P = \pi$; i.e., $\pi(A) = \int \pi(dy)\, P(y,A)$ for all $A \in \mathcal{E}$.

The probability measure π is called the stationary distribution of the process. In the countable case equation (2.12.c) becomes (for $A = \{j\}$)

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij} \,. \tag{2.13}$$

Now let $f:E \to \mathbb{R}$ and assume that $\pi|f| = \int \pi(dy)\ |f(y)| < \infty$. One is often interested in knowing $\pi f = E(f(X))$ for a variety of functions f. For example if $E = \mathbb{R}$ and $f(x) = 1_{\{z:z \leq y\}}(x)$ then $\pi f = P\{X \leq y\}$ and if $f(x) = x^q$ then $\pi f = E(X^q)$. Since the system of equations defined in (2.12.c) (or (2.13)) may be very difficult to solve (for example if E is countably infinite or if E is finite but very large) it may become necessary to estimate πf via simulation. It is the efficient estimation of such quantities that we will concern ourselves with.

Assume now that $X_0 = x_0$ a.s. Let k be some positive integer and for $\nu = 0, 1, \ldots, k$ let $f_\nu:E \to \mathbb{R}$. Let $r_\nu = \pi f_\nu = E(f_\nu(X))$ (assume that $\pi|f_\nu| < \infty$). For each $m \geq 1$ and $\nu = 0, \ldots, k$ define $Y_m(\nu)$ by

$$Y_m(\nu) = \sum_{n=T_{m-1}}^{T_m - 1} f_\nu(X_n) \,. \tag{2.14}$$

Because of the regenerative nature of the Markov chain; i.e., because the $\underset{\sim}{V}_m$'s are i.i.d., $\{(Y_m(0), \ldots, Y_m(k), \tau_m),\ m \geq 1\}$ are i.i.d. random vectors (although for a fixed m, $Y_m(0), \ldots, Y_m(k)$ and τ_m are in general dependent random variables). The following proposition gives an expression for r_ν (in terms of $Y_m(\nu)$ and τ_m) that is useful in simulation. The proof of the proposition may be found in Crane and Iglehart (1975).

(2.15) PROPOSITION. If $\pi|f_\nu| < \infty$, then $r_\nu = \pi f_\nu = E_{x_0}(Y_m(\nu))/E_{x_0}(\tau_m)$.

From now on we will drop the "x_0" in $E_{x_0}(\ )$ with the understanding that $P\{X_0 = x_0\} = 1$. Now set $Z_m(\nu) = Y_m(\nu) - r_\nu \tau_m$. By the previous proposition we have for each $\nu = 0, \ldots, k$, and each $m \geq 1$

$$E(Z_m(\nu)) = 0 \quad . \tag{2.16}$$

For each $i = 0, \ldots, k$ and $j = 0, \ldots, k$ let

$$\sigma_{ij} = E[Z_m(i)\ Z_m(j)] \tag{2.17}$$

and write σ_i^2 for σ_{ii}. Let Σ_k be the symmetric $(k+1)$ by $(k+1)$ dimensional matrix whose (i,j)th entry is σ_{ij}. Σ_k is then the covariance matrix of the $Z_m(\nu)$'s. We assume that each element of Σ_k is finite and that Σ_k is positive definite (in general it is positive semidefinite).

We now state some additional limit theorems for Markov processes from which point estimates and confidence intervals for each r_ν can be obtained. Let

$$\hat{r}_\nu(M) = \sum_{m=1}^{M} Y_m(\nu) / \sum_{m=1}^{M} \tau_m \quad , \tag{2.18}$$

$$\hat{x}_\nu(N) = \frac{1}{N+1} \sum_{n=0}^{N} f_\nu(X_n) \quad , \tag{2.19}$$

for each $\nu = 0, 1, \ldots, k$.

According to the following strong laws of large numbers we can derive strongly consistent point estimates for r_ν using either $\hat{r}_\nu(M)$ or $\hat{x}_\nu(N)$. The proof may be found in Crane and Iglehart (1975).

(2.20) PROPOSITION. If $\pi|f_\nu| < \infty$, then

$$\hat{r}_\nu(M) \to r_\nu \qquad \text{a.s. as } M \to \infty ,$$

$$\hat{x}_\nu(N) \to r_\nu \qquad \text{a.s. as } N \to \infty .$$

Notice that $\hat{r}_\nu(M)$ is an estimator for r_ν based on M cycles of the process and $\hat{x}_\nu(N)$ estimates r_ν based on N transitions of the process.

To derive confidence intervals for the r_ν's we can use a multi-dimensional central limit theorem. If A is a symmetric positive definite $(k+1)$ by $(k+1)$ dimensional matrix let $N(\underset{\sim}{0}, A)$ denote a $(k+1)$ dimensional random vector having a multivariate normal distribution with means 0 and covariance matrix A. For any real number $a \neq 0$ let A/a be the $(k+1)$ by $(k+1)$ dimensional matrix whose (i,j)th entry is A_{ij}/a.

(2.21) PROPOSITION. If $\pi|f_\nu| < \infty$ for each $\nu = 0, 1, \ldots, k$ and if Σ_k is a finite positive definite matrix then

$$(\sqrt{M}\,(\hat{r}_0(M) - r_0), \ldots, \sqrt{M}\,(\hat{r}_k(M) - r_k)) \Rightarrow N(\underset{\sim}{0}, \Sigma_k/E^2(\tau_1)) ,$$

$$(\sqrt{N}\,(\hat{x}_0(N) - r_0), \ldots, \sqrt{N}\,(\hat{x}_k(N) - r_k)) \Rightarrow N(\underset{\sim}{0}, \Sigma_k/E(\tau_1)) \quad .$$

PROOF. Apply the Cramér-Wold device described on page 48 of Billingsley (1968) to the central limit theorem given in equation (5.3) of Crane and Iglehart (1975). □

Now let $\underset{\sim}{\beta}$ be a $(k+1)$ dimensional row vector of real numbers whose νth entry is $\beta(\nu)$. Let $\underset{\sim}{r}$, $\hat{\underset{\sim}{r}}(M)$ and $\hat{\underset{\sim}{x}}(N)$ denote $(k+1)$ dimensional column vectors whose νth entries are r_ν, $\hat{r}_\nu(M)$ and $\hat{x}_\nu(N)$ respectively. If A is a matrix let A' denote the transpose of A; i.e., $A'_{ij} = A_{ji}$. A simple application of the continuous mapping theorem (Theorem (5.1) of Billingsley (1968)), yields;

(2.22) PROPOSITION. Let $\sigma_k^2(\underset{\sim}{\beta}) = \underset{\sim}{\beta}\,\Sigma_k\,\underset{\sim}{\beta}' = \sum_{i=0}^{k}\sum_{j=0}^{k} \beta(i)\,\sigma_{ij}\,\beta(j)$. Under the hypotheses of Proposition (2.21),

$$\frac{\sqrt{M}\,(\underset{\sim}{\beta}\,\hat{\underset{\sim}{r}}(M) - \underset{\sim}{\beta}\underset{\sim}{r})}{\sigma_k(\underset{\sim}{\beta})/E(\tau_1)} \Rightarrow N(0,1) \qquad \text{as} \quad M \to \infty ,$$

and

$$\frac{\sqrt{N}\,(\underset{\sim}{\beta}\,\hat{\underset{\sim}{x}}(N) - \underset{\sim}{\beta}\underset{\sim}{r})}{\sigma_k(\underset{\sim}{\beta})/E(\tau_1)^{1/2}} \Rightarrow N(0,1) \qquad \text{as} \quad N \to \infty .$$

Note that, for example, $\underset{\sim}{\beta r} = \sum_{\nu=0}^{k} \beta(\nu)\, r_\nu$. A central limit theorem for an individual r_ν can be formed using Proposition (2.22) by setting $\beta(\nu) = 1$ and $\beta(i) = 0$ for $i \neq \nu$. In this case the limit theorems become;

$$(2.23) \qquad \frac{\sqrt{M}\,(\hat{r}_\nu(M) - r_\nu)}{\sigma_\nu / E(\tau_1)} \Rightarrow N(0,1) \qquad \text{as } M \to \infty ,$$

$$(2.24) \qquad \frac{\sqrt{N}\,(\hat{x}_\nu(N) - r_\nu)}{\sigma_\nu / E(\tau_1)^{1/2}} \Rightarrow N(0,1) \qquad \text{as } N \to \infty .$$

In order to form confidence intervals for the r_ν's (or linear combinations of the r_ν's) it is necessary to know the covariance terms σ_{ij} as well as $E(\tau_1)$. In a simulation these constants are usually unknown and must be estimated. In addition $\underset{\sim}{\beta}$ may be a fixed, but unknown, vector so it too must be estimated. The following proposition tells us that we may replace these quantities in Proposition (2.22) by any sequence of strongly consistent estimators and not destroy the asymptotic normality.

(2.25) PROPOSITION. Suppose that $\bar{\tau}_1(M) \to E(\tau_1)$ a.s., $\hat{\sigma}_{ij}(M) \to \sigma_{ij}$ a.s. for each i and j, and that $\hat{\beta}(i,M) \to \beta(i)$ a.s. for each i. Let $\hat{\Sigma}_k(M)$ be the matrix whose (i,j)th entry is $\hat{\sigma}_{ij}(M)$, let $\hat{\underset{\sim}{\beta}}(M)$ be the vector whose ith component is $\hat{\beta}(i,M)$ and let $\hat{\sigma}_k(\hat{\underset{\sim}{\beta}},M) = \hat{\underset{\sim}{\beta}}(M)\, \hat{\Sigma}_k(M)\, \hat{\underset{\sim}{\beta}}'(M)$. Then

$$\frac{\sqrt{M}\,(\hat{\underset{\sim}{\beta}}(M)\, \hat{\underset{\sim}{r}}(M) - \hat{\underset{\sim}{\beta}}(M)\underset{\sim}{r})}{\hat{\sigma}_k(\hat{\underset{\sim}{\beta}},M)/\bar{\tau}_1(M)} \Rightarrow N(0,1) \qquad \text{as } M \to \infty .$$

PROOF. By Theorem 4.4 of Billingsley (1968) and Proposition (2.22) we have

$$(\bar{\tau}_1(M),\ \hat{\underset{\sim}{\beta}}(M),\ \hat{\Sigma}_k(M),\ \sqrt{M}\ (\hat{\underset{\sim}{r}}(M) - \underset{\sim}{r}))$$

$$\Rightarrow \quad (E(\tau_1)\ \underset{\sim}{\beta},\ \Sigma_k,\ N(\underset{\sim}{0},\ \Sigma_k/E^2(\tau_1))) \quad .$$

We then have

$$(\hat{\sigma}_k(\hat{\underset{\sim}{\beta}},M)/\bar{\tau}_1(M),\ \sqrt{M}\ (\hat{\underset{\sim}{\beta}}(M)\ \hat{\underset{\sim}{r}}(M) - \hat{\underset{\sim}{\beta}}(M)\underset{\sim}{r}))$$

$$\Rightarrow (\sigma_k(\underset{\sim}{\beta})/E(\tau_1),\ N(0,\ \sigma_k^2(\underset{\sim}{\beta})/E^2(\tau_1)) \quad ,$$

$$\left((\hat{\sigma}_k(\hat{\underset{\sim}{\beta}},M)/\bar{\tau}_1(M), \frac{\sqrt{M}\ (\hat{\underset{\sim}{\beta}}(M)\ \hat{\underset{\sim}{r}}(M) - \hat{\underset{\sim}{\beta}}(M)\underset{\sim}{r})}{\sigma_k(\underset{\sim}{\beta})/E(\tau_1)}\right) \Rightarrow (\sigma_k(\underset{\sim}{\beta})/E(\tau_1),\ N(0,1)) \quad ,$$

and finally

$$\frac{\sqrt{M}\ (\hat{\underset{\sim}{\beta}}(M)\ \hat{\underset{\sim}{r}}(M) - \hat{\underset{\sim}{\beta}}(M)\underset{\sim}{r})}{\hat{\sigma}_k(\hat{\underset{\sim}{\beta}},M)/\bar{\tau}_1(M)} \Rightarrow N(0,1) \quad ,$$

the last three steps all being justified by the continuous mapping theorem. □

3. Variance Reduction Techniques

In this section we apply the results of Section 2 and take further advantage of the structure of Markov processes to obtain variance reductions. Let us now fix a function $f:E \to \mathbb{R}$ and as before let $r = \pi f$. Our goal is to obtain both point estimates and "short" confidence intervals for r. This will be achieved by forming several estimators for r and then taking the (asymptotic) minimum variance linear combination of these estimates which is strongly consistent. In order to form these multiple estimates some additional calculations must be done both before and during the simulation, but hopefully their cost will not be so great as to prohibit the use of this method.

The multiple estimates for r are formed by choosing new functions $f_\nu:E \to \mathbb{R}$ so that $\pi f_\nu = r$ for each value of ν. The values of ν for which this is done will typically be "small" and labeled as $\{0,1,\ldots,k\}$. Once the f_ν's have been computed we simulate the process for, say, M cycles. As in Section 2 let

$$Y_m(\nu) = \sum_{n=T_{m-1}}^{T_m-1} f_\nu(x_n)$$

and define

$$\hat{r}_\nu(M) = \sum_{m=1}^{M} Y_m(\nu) \Big/ \sum_{m=1}^{M} \tau_m \quad .$$

Since, assuming $\pi|f_\nu| < \infty$, $\hat{r}_\nu(M) \to E(Y_m(\nu))/E(\tau_m) = \pi f_\nu = r$, each $\hat{r}_\nu(M)$ is a strongly consistent estimator for r so that we could use any one of them to estimate r. Actually we can do significantly better than that by using $\hat{r}_0(M), \ldots, \hat{r}_k(M)$ simultaneously to estimate r. If we let $\{\beta(\nu):\nu = 0, \ldots, k\}$ be any constants so that

$$\sum_{\nu=0}^{k} \beta(\nu) = 1 \tag{3.1}$$

and then set

$$\hat{r}_\beta(M) = \sum_{\nu=0}^{k} \beta(\nu)\, \hat{r}_\nu(M) \tag{3.2}$$

we have $\hat{r}_\beta(M) \to r$ a.s. as $M \to \infty$. The values of $\beta(\nu)$ are then chosen to minimize the asymptotic variance of $\hat{r}_\beta(M)$. Details of the choice of the $\beta(\nu)$'s will be presented later.

We now turn to the selection of the functions f_ν. In this paper we will concentrate on only one (actually the simplest) way to choose the f_ν's. In a subsequent paper we will study alternate methods of choosing the f_ν's in the case when the state space is finite.

As our current choice of f_ν we let

$$f_\nu = P^\nu f, \qquad \text{for } \nu = 0, \ldots, k. \tag{3.3}$$

Recall that P^ν is the ν-step transition function of the process. If we set $r_\nu = \pi f_\nu$ we must show that $r_\nu = r$ for each value of ν. Assuming that the following steps can be justified (which they will be) we have

$$r_\nu = \pi f_\nu = \pi(P^\nu f)$$

$$= \pi P(P^{\nu-1} f)$$

$$= \pi(P^{\nu-1} f) \qquad \text{(since } \pi P = \pi\text{)}$$

$$= \pi f_{\nu-1} = r_{\nu-1} \ .$$

So all the r_ν's are equal. Noting that $f_0 = P^0 f = f$ we have $\pi f_0 = \pi f = r$, and so $r_\nu = r$ for all ν. The following proposition formalizes this idea by showing that $\pi f_1 = \pi(Pf) = r$. The fact that $\pi f_\nu = r$ for all $\nu \geq 0$ then follows easily by induction.

(3.4) PROPOSITION. If $\pi|f| < \infty$, then $\pi f = \pi(Pf)$.

PROOF. Recall that $\pi f = \int \pi(dy)\ f(y)$. Assume for the moment that $f \geq 0$. We first prove the proposition for functions f which are indicators; i.e., $f(y) = 1_{\{B_j\}}(y)$ for some $B \in \mathcal{E}$,

$$f_1(x) = \int P(x, dy)\ f(y) = \int P(x, dy)\ 1_{\{B\}}(y)$$

$$= \int_B P(x, dy) = P(x,B) \quad .$$

We therefore have

$$\pi f_1 = \int \pi(dx)\ P(x,B)$$

$$= \pi(B) \qquad \text{by (2.12.c)}$$

$$= \int_B \pi(dy)$$

$$= \int \pi(dy)\ 1_{\{B\}}(y) = \pi f \quad .$$

Since the proposition is true for indicator functions it is true for simple functions; i.e., functions which are finite linear combinations of indicators. Now let $f^{(n)}$ be a sequence of simple functions increasing to f. By the monotone convergence theorem

$$\pi f = \lim_{n \to \infty} \pi f^{(n)} \quad . \tag{3.5}$$

If we let $f_1^{(n)} = Pf^{(n)}$, then

$$Pf(x) = \int P(x, dy)\ f(y)$$

$$= \lim_{n \to \infty} \int P(x, dy)\ f^{(n)}(y) \ ,$$

$$Pf(x) = \lim_{n \to \infty} f_1^{(n)}(x) \ , \tag{3.6}$$

again by the monotone convergence theorem. Furthermore for each x the convergence in (3.6) is monotone. Now since the proposition is true for simple functions, we have

$$\pi f^{(n)} = \pi f_1^{(n)} \quad . \tag{3.7}$$

The left hand side of (3.7) converges to πf by (3.5) so that

$$\begin{aligned} \pi f &= \lim_{n \to \infty} \pi f_1^{(n)} \\ &= \lim_{n \to \infty} \int \pi(dx)\ f_1^{(n)}(x) \\ &= \int \pi(dx)\ \left(\lim_{n \to \infty} f_1^{(n)}(x) \right) \\ &= \int \pi(dx)\ Pf(x) \qquad \text{by (3.6)} \\ &= \pi(Pf) \quad . \end{aligned}$$

The last interchange of limit and integral is once again justified by the monotone convergence theorem. Finally, since $\pi|f| < \infty$, the result is true for a general f by writing $f(x) = f^+(x) - f^-(x)$ where $f^+(x) = \max(0, f(x))$ and $f^-(x) = -\min(0, f(x))$ and applying the result to both f^+ and f^-. □

We now return to the minimization of the asymptotic variance of the estimator $\hat{r}_\beta(M)$ defined in equation (3.2). By Proposition (2.21) we have the following central limit theorem;

$$\frac{\sqrt{M}\left(\sum_{\nu=0}^{k}(\beta(\nu)\,\hat{r}_\nu(M) - \beta(\nu)r)\right)}{\sigma_k(\underset{\sim}{\beta})/E(\tau_1)} \Rightarrow N(0,1) \qquad \text{as } M \to \infty$$

where $\sigma_k(\underset{\sim}{\beta})$ is defined in (2.22). If the $\beta(\nu)$'s sum to one (equation (3.1)) we can rewrite this as

$$\frac{\sqrt{M}\,(\hat{r}_\beta(M) - r)}{\sigma_k(\underset{\sim}{\beta})/E(\tau_1)} \Rightarrow N(0,1) \tag{3.8}$$

so that we can form confidence intervals for r based on the estimator $\hat{r}_\beta(M)$. Since we are free to pick $\underset{\sim}{\beta}$ in any way we please (subject to (3.1)), we select $\underset{\sim}{\beta} = \underset{\sim}{\beta}^*$ where $\underset{\sim}{\beta}^*$ minimizes the variance term $\sigma_k^2(\underset{\sim}{\beta})$. This will produce the smallest possible confidence interval for r. Note that there is no reason to restrict $\underset{\sim}{\beta}$ to be nonnegative; i.e., $\hat{r}_\beta(M)$ need not be a convex combination of the $r_\nu(M)$'s. To minimize the variance we must solve the nonlinear programming problem;

$$\text{minimize} \qquad \sigma_k^2(\underset{\sim}{\beta})$$

$$\text{subject to} \qquad \sum_{\nu=0}^{k} \beta(\nu) = 1 \quad .$$

We can write this in matrix notation by letting $\underset{\sim}{e}$ denote the $(k+1)$ dimensional row vector each of whose components is 1. Our problem then becomes

(3.9) $$\text{minimize} \qquad \underset{\sim}{\beta}\, \Sigma_k\, \underset{\sim}{\beta}'$$

$$\text{subject to} \qquad \underset{\sim}{e\beta}' = 1 \quad .$$

It is straightforward (using Lagrange multipliers) to show that

(3.10) $$\underset{\sim}{\beta}^* = \underset{\sim}{e}\, \Sigma^{-1} / \underset{\sim}{e}\, \Sigma^{-1}\, \underset{\sim}{e}' \quad ,$$

(3.11) $$\sigma_k^2(\underset{\sim}{\beta}^*) = 1/\underset{\sim}{e}\, \Sigma^{-1}\, \underset{\sim}{e}' \quad .$$

This selection of multipliers is similar to those given on page 19 of Hammersley and Handscomb (1964), Lavenberg (1974), and Lavenberg and Shedler (1975). We have dealt here with forming short confidence intervals based on a run of length M cycles. The same technique applies to a run of N transitions. In this case we must still solve the nonlinear program given in (3.9) because the variance terms in the central limit theorems (2.21) differ only by a constant multiple. Equations (3.10) and (3.11) are therefore still valid in this case. If we let

(3.12) $$\hat{x}_\beta(N) = \sum_{\nu=0}^{k} \beta(\nu)\, \hat{x}_\nu(N)$$

we then have the two central limit theorems;

$$(3.13) \qquad \frac{\sqrt{M}\,(\hat{r}_{\beta^*}(M) - r)}{\sigma_k(\underset{\sim}{\beta}^*)/E(\tau_1)} \Rightarrow N(0,1) \ ,$$

$$(3.14) \qquad \frac{\sqrt{N}\,(\hat{x}_{\beta^*}(N) - r)}{\sigma_k(\underset{\sim}{\beta}^*)/E(\tau_1)^{1/2}} \Rightarrow N(0,1) \ .$$

Since the covariance matrix is in general unknown it becomes necessary to estimate Σ_k. If $\hat{\Sigma}_k(M)$ is any estimator such that $\hat{\Sigma}_k(M) \to \Sigma_k$ a.s. as $M \to \infty$ then $\hat{\Sigma}_k(M)^{-1} \to \Sigma_k^{-1}$. Letting

$$(3.15) \qquad \hat{\underset{\sim}{\beta}}^*(M) = \underset{\sim}{e}\,\hat{\Sigma}_k^{-1}(M)/\underset{\sim}{e}\,\hat{\Sigma}_k^{-1}(M)\,\underset{\sim}{e}' \ ,$$

it is clear that $\hat{\underset{\sim}{\beta}}^*(M) \to \underset{\sim}{\beta}^*$ a.s. as $M \to \infty$. Applying this result with Proposition (2.25) we can maintain the asymptotic normality even when $\underset{\sim}{\beta}^*$ must be estimated. Letting

$$(3.16) \qquad \hat{r}_{\hat{\beta}^*}(M) = \sum_{\nu=0}^{k} \hat{\beta}^*(\nu,M)\,\hat{r}_\nu(M) \ ,$$

we then have the central limit theorem;

$$(3.17) \qquad \frac{\sqrt{M}\,(\hat{r}_{\hat{\beta}^*}(M) - r)}{\hat{\sigma}(\hat{\underset{\sim}{\beta}}^*(M))/\bar{\tau}_1(M)^{1/2}} \Rightarrow N(0,1) \ ,$$

where $\hat{\sigma}(\hat{\underset{\sim}{\beta}}^*(M))$ is defined in (2.25) and $\bar{\tau}_1(M)$ is any sequence of numbers such that $\bar{\tau}_1(M) \to E(\tau_1)$ a.s. A corresponding central limit theorem exists for the $\hat{x}_\nu(N)$'s. Because this method combines several

different estimates of the same quantity we call it the "method of multiple estimates".

In order to apply this method, the functions f_ν must be computed (usually before the start of the simulation). For computational efficiency f_ν can be defined recursively by $f_0 = f$ and $f_\nu = Pf_{\nu-1}$ for $\nu \geq 1$. This avoids having to compute the ν-step transition function P^ν, a potentially great computational savings. If the state space is finite and the transition matrix is sparse the work involved in calculating f_ν for a few values of ν may not be too great. If the f_ν's are computed before the start of the simulation we must be able to store them; this may be a considerable problem if the state space is very large.

We note that to form the estimates $\hat{x}_\nu(N)$ (or $\hat{r}_\nu(M)$) we must evaluate $f_\nu(X_n)$ for each value of ν and each transition n. This will tend to increase the amount of time needed for each transition simulated. However, if the variance reduction obtained is sufficiently large the potential savings in the number of transitions that need to be simulated will more than offset the extra work per transition. This will be discussed in greater detail in Section 4. We also note that additional work must be done at the end of each cycle to update the estimates of the covariance matrix Σ_k (using no variance reducing technique we need only update σ_0^2). However if k is small and if cycles tend to be long this should not have a substantial impact on the CPU time.

It should be mentioned that if one has a choice of more than one return state the variance reductions that are obtained using this method are (in theory) independent of the return state. The reader should consult

either Chung (1967) or Crane and Iglehart (1975) for this point. For practical reasons it is recommended that, if possible, the return state be chosen so that cycles are not excessively long.

We now examine the selection of the parameter k. As the value of k increases $\sigma_k^2(\underset{\sim}{\beta}^*)$ decreases. This is because the kth minimization problem is the same as the $(k+1)$st problem which has the additional constraint that $\beta(k+1) = 0$. This means that as we do more computation we can get increasingly accurate estimates of r.

If the state space is finite we have

$$f_k(i) = \sum_{j \in E} P_{ij}^k f(j) \to \sum_{j \in E} \pi_j f(j) = r \qquad \text{as} \quad k \to \infty \quad ,$$

so that for large values of k, f_k will be a smoother function than f. The point estimate $\hat{x}_k(N)$ is then the average of $N+1$ terms each of which is close to r so that, for large k, $\hat{x}_k(N)$ will have a smaller variance than $\hat{x}_0(N)$. In fact it can be shown that $\sigma_k^2 \to 0$ as $k \to \infty$. By placing all weight on, $\hat{x}_k(N)$; i.e., by setting $\beta(k) = 1$ and $\beta(\nu) = 0$ for $\nu \neq k$, we see that $\sigma_k^2(\underset{\sim}{\beta}^*) \leq \sigma_k^2((0, 0, \ldots, 1)) = \sigma_k^2 \to 0$. Furthermore, since in the case of a finite state space $f_k(i)$ converges exponentially to r; i.e., there exist constants $c_1 \geq 0$, $1 > c_2 \geq 0$ such that $|f_k(i) - r| \leq c_1 c_2^k$, the rate at which $\sigma_k^2(\underset{\sim}{\beta}^*)$ converges to 0 is also exponential (see Doob (1953)).

For many types of Markov chains we can expect substantial variance reductions even when k is relatively small (say 2 or 3). For countable E we have

$$(3.18) \qquad f_k(i) = \sum_{j=0}^{\infty} P_{ij}^k f(j) = E(f(X_{n+k}) | X_n = i) .$$

Thus if the Markov chain makes transitions only to "neighboring" states and if $f(j)$ is close to $f(i)$ for j "near" to i it can be seen from equation (3.18) that for small k, $f_k(i)$ and $f(i)$ should be nearly the same. This means that $\hat{x}_k(N)$ and $\hat{x}_0(N)$ will be highly correlated, a condition that generally results in good variance reduction. Many queueing networks exhibit this special type of structure.

Ideally one would like to be able to select the "optimal" value of k in the sense that for a given computer budget we would like to pick the value of k which gives us the smallest confidence intervals for r (part of the budget must be allocated to the calculation of the f_ν's). To perform such an optimization one would have to know (or have estimates for) $\sigma_\nu^2(\underset{\sim}{\beta}^*)$ for each $\nu \geq 0$. These quantities are in general unknown and even to estimate them would require calculating the f_ν's and then simulating the Markov process for a number of cycles. The disadvantage of such a procedure is that one may invest considerable time in the computation of f_ν only to find that the calculation was wasted; i.e., the variance reduction obtained by using f_ν is not sufficiently great to justify the added cost of the calculations.

One possible approach to this problem is a sequential one. Suppose one has calculated $f_0, f_1, \ldots, f_k$ and has estimates for $\sigma_0^2, \sigma_1^2(\underset{\sim}{\beta}^*), \ldots, \sigma_k^2(\underset{\sim}{\beta}^*)$. These estimates could then be plotted and an estimate for $\sigma_{k+1}^2(\underset{\sim}{\beta}^*)$ could be obtained by extrapolation. It could then be decided (based on these estimates) whether or not f_{k+1} should

be calculated. The success of such a procedure depends, of course, on how accurate our estimates for $\{\sigma_\nu^2(\underset{\sim}{\beta}^*): 0 \leq \nu \leq k\}$ are as well as the validity of the extrapolation.

We feel that the inability to predict the variance reductions in advance is the major drawback of the method (this is also a drawback in other more standard control variable methods). The simulator must be very careful to apply the method only in those cases where it is efficient to do so. Generally speaking the success of this technique depends on one's ability to efficiently compute and store the functions f_ν.

The method of multiple estimates can be extended to certain types of continuous time processes such as continuous time Markov chains and semi-Markov processes. The basic idea here is to simulate the imbedded discrete time Markov chain of the process and hold the process in a state for a deterministic amount of time equal to the expected holding time of the state. We thus transform the continuous time process into a discrete time process. Once the appropriate function f for the discrete time process is formed the method proceeds exactly as before. The reader is referred to Hordijk, Iglehart and Schassberger (1976) for a more detailed account of this transformation. Several of the examples in Section 4 were treated in this manner.

Since it is often inconvenient to actually simulate the imbedded discrete time Markov chain we present an analysis that shows how the method can be applied to continuous time Markov chains which are simulated in continuous time. Let $\{X_t, t \geq 0\}$ be a continuous time Markov chain with transition matrix $P_{ij}(t) = P\{X_t = j | X_0 = i\}$. Let

(3.19) $$q_{ij} = \frac{d}{dt} P_{ij}(t)\Big|_{t=0} ,$$

$Q = \{q_{ij}: i,j \in E\}$, and $q_i = -q_{ii}$ (assume $0 < q_i < \infty$). We also assume that the process is positive recurrent and irreducible. Let $\pi = \{\pi_i: i \in E\}$ be the stationary distribution of the process. It is then known (see Çinlar (1975)) that

(3.20) $$\pi Q = 0 \quad .$$

Let $f: E \to \mathbf{R}$, $r = \pi f$, and assume that $\pi|f| < \infty$ as before. We again are interested in finding functions f_ν so that

$$r = \pi f_\nu \left(= \lim_{T \to \infty} \frac{1}{T} \int_0^T f_\nu(X_s)ds\right)$$

for each ν.

Writing (3.20) out we see that

$$0 = \sum_{i \in E} \pi_i \, q_{ij}$$

$$= \sum_{i \neq j} \pi_i \, q_{ij} - \pi_j \, q_j$$

which reduces to

(3.21) $$\pi_j = \sum_{i \neq j} \pi_i \, q_{ij}/q_j \quad .$$

Define the matrix $A = \{a_{ij}: i,j \in E\}$ by

$$a_{ij} = \begin{cases} 0 & \text{if } i = j \ , \\ q_{ij}/q_j & \text{if } i \neq j \ . \end{cases} \tag{3.22}$$

Equation (3.21) then becomes (in matrix notation)

$$\pi = \pi A \ . \tag{3.23}$$

We therefore have $\pi f = \pi A f$ so that if we let $f_1 = Af$ we have $\pi f = \pi f_1$. By induction if we let

$$f_\nu = A^\nu f \qquad \text{for } \nu \geq 0 \tag{3.24}$$

then $\pi f_\nu = \pi f$ for all $\nu \geq 0$ (define $A^0 = I$, the identity, so that $f_0 = f$ as before).

Let us assume (without loss of generality) that we now pick 0 as our return state. For $m \geq 1$ let ρ_m be the length of the mth visit to state 0. Define the return times T_m by

(3.25) **DEFINITION.** $T_0 = 0$,

$$T_m = \inf\{s > T_{m-1} + \rho_m : X_s = 0\} \ , \qquad \text{for } m \geq 1 \ .$$

Let $Y_m(\nu)$ be defined by

$$Y_m(\nu) = \int_{T_{m-1}}^{T_m} f_\nu(X_s)ds \quad . \tag{3.26}$$

Since we still have $r = E(Y_m(\nu))/E(\tau_m)$, we can now proceed exactly as in the discrete time case. The important thing to notice is that we only need the infinitesimal matrix Q to form the functions f_ν. The actual simulation of the process can take place in the most straightforward manner.

If we convert the process into discrete time using the methods of Hordijk et al. (1976) and then apply the method of multiple estimates we generate a sequence of functions which will be denoted by f'_ν. It can then be shown that

$$f'_\nu(i) = f_\nu(i)/q_i \tag{3.27}$$

where $f_\nu(i)$ is defined in (3.24). Recall that q_i^{-1} is the mean of the exponentially distributed holding time of state i.

To apply the method to semi-Markov processes simulated in continuous time we cannot use (3.24) to calculate our functions f_ν. Instead we form the functions f'_ν by using the discrete time methods of Hordijk, et al. (1976) and then if q_i^{-1} is the mean holding time of state i we let

$$f_\nu(i) = f'_\nu(i)\, q_i \quad . \tag{3.28}$$

It can then be shown that again $\pi f_\nu = \pi f$ for all $\nu \geq 0$, where π is now the stationary distribution of the semi-Markov process. We then define T_m and $Y_m(\nu)$ as in (3.25) and (3.26) and proceed as before.

As a further extension of the method we can drop both the Markovian and regenerative assumptions on $\{X_n\}$ but to do so we need to add the hypotheses that the process is stationary and satisfies an asymptotic independence condition known as φ- mixing (see Billingsley (1968)). Let $\{X_n\}$ be a strictly stationary process with state space E. For integers $a \leq b$ let $\mathcal{M}_a^b$ be the sigma field generated by $X_a, \ldots, X_b$ (with the obvious extensions for $a = -\infty$ and $b = +\infty$).

(3.29) DEFINITION. We say that $\{X_n\}$ is φ-mixing if there exists a sequence of numbers $\{\varphi(n)\}$ such that for each k $(-\infty < k < \infty)$ and $n \geq 1$, $E_1 \in \mathcal{M}_{-\infty}^k$, $E_2 \in \mathcal{M}_{k+n}^{\infty}$,

$$|P(E_1 \cap E_2) - P(E_1)\ P(E_2)| \leq \varphi(n)\ P(E_1) \quad .$$

Assuming that $P(E_1) > 0$ we can divide both sides of the inequality in (3.29) by $P(E_1)$ to obtain the equivalent condition

$$|P(E_2|E_1) - P(E_2)| \leq \varphi(n) \quad . \tag{3.30}$$

If $\varphi(n) \to 0$ as $n \to \infty$ (3.30) says that events in the distant future are nearly independent of the past. Let $f:E \to \mathbb{R}$ and let $r = E(f(X_n))$. We are now interested in estimating r. Define

(3.31) $$Y_n(\nu) = E(f(X_{n+\nu}|\mathcal{M}_{n-a}^n)$$

$$= E(f(X_{n+\nu})|X_n, X_{n-1}, \ldots, X_{n-a}) ,$$

for some fixed finite integer a. It can then be shown that the process $\{Y_n(\nu)\}$ is also stationary and φ-mixing (with a different φ). Furthermore if $E(|f(X_n)|) < \infty$, then

(3.32) $$r_\nu = E(Y_n(\nu)) = E(E(f(X_{n+\nu})|\mathcal{M}_{n-a}^n))$$

$$= E(f(X_{n+\nu}))$$

$$= r \qquad \text{(by stationarity)} .$$

We can again try to apply the method of multiple estimates to estimate r. To do so let

(3.33) $$\hat{x}_\nu(N) = \frac{1}{N+1} \sum_{n=0}^{N} Y_n(\nu)$$

(3.34) $$Z_n(\nu) = Y_n(\nu) - r .$$

We can now use the multidimensional central limit theorem for φ-mixing processes given in Billingsley (1968).

(3.35) PROPOSITION. If $\sum_{n=0}^{\infty} \varphi(n)^{1/2} < \infty$ and $E(|f(X_n)|) < \infty$, then

$$(\sqrt{N}\ (\hat{x}_0(N) - r),\ \ldots,\ \sqrt{N}\ (\hat{X}_k(N) - r)) \Rightarrow N(\underset{\sim}{0},\ V_k)$$

where V_k is assumed to be a symmetric $(k+1)$ by $(k+1)$ dimensional positive definite matrix with finite entries v_{ij} defined by

$$(3.36) \qquad v_{ij} = E(Z_0(i)\ Z_0(j)) + \sum_{n=1}^{\infty} E(Z_0(i)\ Z_n(j)) + \sum_{n=1}^{\infty} E(Z_n(i)\ Z_0(j)) .$$

We can now proceed as before. The terms v_{ij} are usually unknown and must be estimated, frequently a formidable task. In addition the calculation of $Y_n(v)$ may be very difficult. It will depend on the structure of the underlying stochastic process $\{X_n\}$ and unlike in the Markov case, no neat computational formulas (like $f_v = P^v f$) can be given for $E(f(X_{n+v})|X_n, X_{n-1}, \ldots, X_{n-a})$. Because of these difficulties no numerical examples of this type were tried, although further research in this direction is planned.

4. Examples

To find out how well the method performs, four test problems were selected. The four problems all come from the area of queueing theory. They are the queue length process in a finite capacity M/M/1 queue, the queue length process in the repairman problem with spares, and the waiting time processes in both the M/M/1 and M/M/2 queues. These processes were chosen because analytic results are readily available, thereby making a comparison between analytic and simulation results possible. Despite their simplicity they are by no means "easy" processes to simulate, particularly the heavily loaded queues which require very long run lengths (see Table 1) to get good simulation estimates. For all four types of processes substantial variance reductions have been realized (although in the M/M/2 queue with $\rho = .9$ the reduction in variance is not enough to justify the use of the method). While it is difficult to predict the variance reductions that will be obtained for a particular stochastic process, it is felt that this method shows a great deal of promise and deserves serious consideration when planning a simulation experiment. The remainder of this section will be devoted to a more detailed description of the examples and a presentation of their numerical results.

(4.1) EXAMPLE. Finite Capacity M/M/1 Queue

Let $\{X(t),\ t \geq 0\}$ be a birth and death process with parameters

$$\lambda_i = \begin{cases} \lambda, & i = 0, 1, \ldots, M-1 \\ 0, & i \geq M \end{cases}$$

$$\mu_i = \mu, \qquad i = 1, 2, \ldots, M$$

where M is a finite positive integer and $0 < \lambda, \mu < \infty$. The process $\{X(t),\ t \geq 0\}$ is then the queue length process for an M/M/1 queue with finite queue capacity M. In this system customers arrive according to a Poisson process with rate λ, however any customer arriving when there are M customers already in the system is denied entrance to the queue and departs never to return. The i.i.d. service times have an exponential distribution with mean μ^{-1}. The state space is $E = \{0, 1, \ldots, M\}$. We investigate the variance reductions for three different functions f; $f(i) = i$, $f(i) = i^2$ and $f(i) = 1_{\{0\}}(i)$. The corresponding r's for these functions are $E(X)$, $E(X^2)$, and $P\{X = 0\}$ respectively. By applying the discrete time methods of Hordijk, et al., (1976) we can solve systems of linear equations to find r, Σ_k, $\underset{\sim}{\beta}^*$ and $\sigma_k^2(\underset{\sim}{\beta}^*)$. M has been chosen to be 14 so that these systems of equations may be easily solved for a wide variety of parameter values. Let $\rho = \lambda/\mu$. Because the state space is finite the Markov chain is positive recurrent for all values of ρ (unlike the infinite capacity queue which requires $\rho < 1$). For small

values of ρ the capacity constraint plays little role so that behavior of the finite capacity and the infinite capacity queues should be nearly the same. To help the reader better assess the effect of the capacity limitation we note that for the infinite capacity M/M/1 queue $E(X) = \rho/(1-\rho)$, $E(X^2) = \rho(1+\rho)/(1-\rho)^2$ and $P(X = 0) = 1-\rho$.

For the numerical results that follow we have slightly modified the definition of $\sigma_k^2(\underset{\sim}{\beta}^*)$ to make it the asymptotic variance term in the central limit theorem for $\hat{x}_{\beta^*}(N)$. It therefore takes into account all constant multiples such as $E(\tau_1)^{1/2}$. Let $R_k^2 = \sigma_k^2(\underset{\sim}{\beta}^*)/\sigma_0^2$. To obtain confidence intervals of equal length, if we use the estimator $\hat{x}_{\beta^*}$ we need only simulate R_k^2 times as many transitions as would be needed using no variance reduction technique (that is if we used just the regular point estimate $\hat{x}_0$). For a fixed (large) number of simulated transitions, N, the length of the confidence interval for r using $\hat{x}_{\beta^*}(N)$ divided by the length of the confidence interval for r using $\hat{x}_0(N)$ is $R_k = \sigma_k(\underset{\sim}{\beta}^*)/\sigma_0$. R_k^2 and R_k are the usual efficiency measures of a variance reduction technique.

Tables 2, 3 and 4 list ρ, r, σ_0^2, R_k^2, and R_k for $k = 1, 2, 3$. For each k, R_k is listed directly below R_k^2. As an example, in Table 2, for $\rho = .5$ we see that to obtain confidence intervals for $E(X) = .9995$ of equal length we need only simulate 5.24% as many transitions using $\hat{x}_{\beta^*}$ (with $k = 3$) than using just $\hat{x}_0$. For the same number of transitions simulated the ratio of the lengths of confidence intervals is .2288. Notice that the choice of the function f influences the variance reductions as does the value of ρ. Generally speaking as the traffic intensity, ρ, increases the variance reduction obtained decreases.

TABLE 2

Calculated Variance Reductions for Finite Capacity M/M/1 Queue: r = E(X)

ρ	E(X)	σ_0^2	R_1^2 R_1	R_2^2 R_2	R_3^2 R_3
.10	.1111	.0244	.0454 .2136	.0045 .0674	.0005 .0213
.20	.2500	.2812	.0926 .3043	.0185 .1361	.0037 .0609
.30	.4286	1.469	.1413 .3759	.0424 .2058	.0127 .1126
.40	.6667	5.888	.1905 .4364	.0756 .2749	.0297 .1725
.50	.9995	21.59	.2341 .4838	.1121 .3347	.0524 .2288
.60	1.493	76.57	.2601 .5100	.1352 .3677	.0681 .2610
.70	2.262	250.9	.2884 .5370	.1395 .3734	.0683 .2613
.80	3.453	670.2	.4094 .6399	.1623 .4028	.0692 .2631
.90	5.111	1262	.6050 .7778	.2659 .5156	.1148 .3388
.95	6.052	1476	.6880 .8294	.3425 .5852	.1607 .4009
.99	6.813	1548	.7404 .8605	.4056 .6369	.2047 .4525

TABLE 3

Calculated Variance Reductions for Finite Capacity M/M/1 Queue: $r = E(X^2)$

ρ	$E(X^2)$	σ_0^2	R_1^2 / R_1	R_2^2 / R_2	R_3^2 / R_3
.10	.1358	.1241	.0271 .1647	.0044 .0664	.0004 .0210
.20	.3750	2.508	.0442 .2103	.0130 .1139	.0026 .0509
.30	.7959	22.64	.0564 .2376	.0222 .1488	.0065 .0808
.40	1.555	156.7	.0627 .2505	.0285 .1688	.0114 .1066
.50	2.992	974	.0628 .2507	.0320 .1790	.0143 .1194
.60	5.873	5477	.1053 .3245	.0316 .1778	.0120 .1094
.70	11.82	25,787	.2681 .5178	.0775 .2784	.0276 .1662
.80	23.42	90,634	.4857 .6969	.1969 .4438	.0905 .3001
.90	42.66	211,336	.6512 .8070	.3433 .5859	.1870 .4325
.95	54.75	270,711	.7078 .8413	.4089 .6395	.2368 .4865
.99	65.06	303,811	.7425 .8617	.4545 .6742	.2737 .5232

TABLE 4

Calculated Variance Reductions for Finite Capacity M/M/1 Queue: r = P(X = 0)

ρ	P(X = 0)	σ_0^2	R_1^2 / R_1	R_2^2 / R_2	R_3^2 / R_3
.10	.9000	.0073	.1000 .3162	.0100 .1000	.0010 .0316
.20	.8000	.0083	.2000 .4472	.0400 .2000	.0080 .0894
.30	.7000	.0132	.3000 .5477	.0900 .3000	.0270 .1643
.40	.6000	.3657	.4000 .6324	.1599 .3999	.0639 .2528
.50	.5000	.6659	.4995 .7067	.2492 .4992	.1242 .3524
.60	.4002	1.069	.5961 .7721	.3541 .5951	.2093 .4575
.70	.3014	1.52	.6830 .8265	.4627 .6802	.3101 .5568
.80	.2073	1.811	.7495 .8657	.5543 .7445	.4033 .6350
.90	.1259	1.619	.7884 .8879	.6113 .7818	.4645 .6816
.95	.0932	1.338	.7975 .8930	.6249 .7905	.4796 .6925
.99	.0715	1.076	.8003 .8946	.6292 .7932	.4843 .6959

(4.2) EXAMPLE. The Repairman Problem with Spares.

We next consider a repairman problem with n operating units and m spares. Each of the operating units fails with rate λ. Upon failure the unit enters a queue to obtain service from one of s repairmen. The failed unit is replaced by a spare (if any are available). The distribution of the i.i.d. repair times is exponential with mean μ^{-1} for each repairman. When a unit is repaired it enters the pool of spares unless there are fewer than n units in operation in which case the repaired unit immediately becomes operational. If we let $X(t)$ denote the number of units in service and in the queue for service then $\{X(t), t \geq 0\}$ is a birth and death process with parameters;

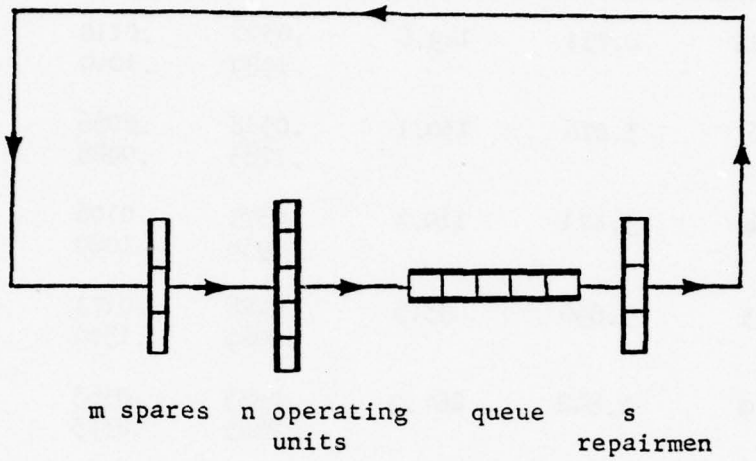
$$\lambda_i = \begin{cases} n\lambda, & 0 \leq i \leq m \\ (n+m-i)\lambda, & m < i \leq n+m \end{cases}$$

$$\mu_i = \begin{cases} i\mu, & 1 \leq i \leq s \\ s\mu, & s < i \leq m+n. \end{cases}$$

The process can be modelled by the simple closed queueing network pictured in Figure 1.

For our functions f we choose $f(i) = i$ and $f(i) = i^2$ so that $r = E(X)$ and $E(X^2)$ respectively. In our example we let $m = 4$, $n = 10$ and $\lambda = 1$. The parameters s and μ are then allowed to vary. Again the state space is $E = \{0, 1, \ldots, 14\}$. It should be emphasized that

FIGURE 1



the figures in Tables 5 and 6 are actual calculations of variance reductions and not estimates derived from simulations. The tables are grouped according to the value $s\mu$, which is the maximum service rate for the repair facility.

Notice that for both functions and all combinations of parameter values substantial variance reductions are obtained. The variance reductions are approximately the same for all sets of parameters, a situation not found in the truncated $M/M/1$ queue. Note, however, that by setting $s = n = 1$ and $m = M-1$ we obtain the same birth and death parameters as in the $M/M/1$ queue with capacity M. The appearance of uniformity in variance reductions is therefore due to the fact that in the range of parameter values chosen the process exhibits a more "stable" type of behavior than the $M/M/1$ queue does.

TABLE 5

Calculated Variance Reductions for
Repairman Problem: $r = E(X)$

(s,μ)	$E(X)$	σ_0^2	R_1^2 R_1	R_2^2 R_2	R_3^2 R_3
1,12	2.751	149.6	.0396 .1989	.0110 .1049	.0058 .0764
2,6	3.078	130.1	.0318 .1783	.0086 .0928	.0071 .0843
3,4	3.471	110.4	.0375 .1936	.0106 .1029	.0105 .1023
4,3	3.890	93.3	.0428 .2069	.0171 .1309	.0171 .1307
1,9	4.842	267.9	.0833 .2885	.0563 .2373	.0321 .1791
2,4.5	5.025	228.0	.0719 .2681	.0501 .2239	.0323 .1797
2,3	5.253	186.9	.0640 .2529	.0434 .2083	.0282 .1678
4,2.25	5.510	150.6	.0590 .2428	.0390 .1975	.0287 .1694
1,6	7.930	155.9	.1766 .4203	.0643 .2537	.0284 .1685
2,3	7.951	148.0	.1634 .4043	.0562 .2371	.0243 .1557
3,2	7.984	137.7	.1475 .3841	.0470 .2167	.0195 .1396
4,1.5	8.030	125.6	.1308 .3617	.0393 .1982	.0184 .1356
1,3	10.999	32.89	.2171 .4659	.0682 .2611	.0223 .1494
2,1.5	11.000	32.88	.2169 .4658	.0680 .2608	.0222 .1489
3,1	11.000	32.86	.2166 .4654	.0677 .2603	.0220 .1482
4,.75	11.000	32.81	.2160 .4648	.0672 .2592	.0216 .1468

TABLE 6

Calculated Variance Reductions for

Repairman Problem: $r = E(X^2)$

(s,μ)	$E(X^2)$	σ_0^2	R_1^2	R_2^2	R_3^2
			R_1	R_2	R_3
1,12	13.46	9,610	.0836	.0156	.0067
			.2892	.1251	.0819
2,6	15.06	9,377	.0674	.0172	.0083
			.2595	.1313	.0911
3,4	17.28	9,009	.0532	.0177	.0104
			.2306	.1329	.1018
4,3	20.01	8,568	.0445	.0217	.0108
			.2109	.1473	.1039
1,9	31.66	28,346	.1659	.0351	.0134
			.4073	.1874	.1157
2,4.5	32.85	26,287	.1420	.0290	.0117
			.3768	.1702	.1083
3,3	34.51	23,787	.1160	.0231	.0108
			.3406	.1521	.1041
4,2.25	36.59	21,153	.0928	.0199	.0113
			.3046	.1411	.1061
1,6	69.25	33,944	.1897	.0629	.0248
			.4356	.2509	.1573
2,3	69.43	33,120	.1804	.0563	.0208
			.4247	.2373	.1443
3,2	69.74	31,904	.1676	.0481	.0165
			.4094	.2193	.1285
4,1.5	70.21	30,281	.1523	.0397	.0129
			.3902	.1992	.1134
1,3	123.99	14,456	.2046	.0594	.0178
			.4523	.2437	.1334
2,1.5	123.99	14,453	.2045	.0594	.0177
			.4522	.2437	.1332
3,1	124.00	14,449	.2044	.0592	.0176
			.4521	.2434	.1328
4,.75	124.00	14,437	.2040	.0589	.0174
			.4517	.2427	.1320

(4.3) EXAMPLE. Waiting Time Process in an M/M/1 Queue.

We now turn to an example of a regenerative Markov process with an uncountable state space $E = [0,\infty)$. We let W_n and S_n be the waiting time and service time respectively of the nth customer in a single server queue. Let A_{n+1} be the time between the arrival of the nth and (n+1)st customer. We assume that $\{S_n, n \geq 0\}$ are i.i.d. with $E(S_n) = \mu^{-1}$ and that $\{A_n, n \geq 1\}$ are i.i.d. with $E(A_n) = \lambda^{-1}$. Let the traffic intensity ρ be defined by $\rho = \lambda/\mu$. We assume that customer number 0 arrives at time 0 to an empty system. Let $X_n = S_{n-1} - A_n$ for $n \geq 1$. The waiting time process $\{W_n, n \geq 0\}$ can be defined recursively by

$$W_0 = 0$$

(4.4)

$$W_n = (W_{n-1} + X_n)^+ , \qquad n \geq 1.$$

It is known that if $\rho < 1$ there exists an infinite number of indices n such that $W_n = 0$ (and the expected time between any two consecutive indices is finite). Thus we choose $x_0 = 0$ as our return state and regenerations occur whenever a customer arrives to an empty queue. For $\rho < 1$ we therefore know that there exists a random variable W such that $W_n \Rightarrow W$. For more details on this queue, which is commonly called the GI/G/1 queue, see Iglehart (1971). We shall be interested in estimating $E(W)$, which is finite if $E(S_n^2) < \infty$. The appropriate function f is then $f(x) = x$. In order to calculate $f_\nu = P^\nu f$ we need to find the transition function of the process. We illustrate this for the M/M/1

queue; i.e., for a queue with exponentially distributed service times with mean μ^{-1} and Poisson arrivals with rate λ. For M/M/1 the calculations are straightforward. The approach generalizes easily to the GI/G/1 queue, although one's ability to carry out the computations in practice depends on the distributions of the service and interarrival times. Numerical integration techniques may be of use here although, in theory, one must calculate $P^{\nu} f(x)$ for all values of $x \geq 0$.

For the M/M/1 queue it is easy to show that

$$P\{X_n \leq y\} = \begin{cases} \frac{\mu}{\lambda+\mu} e^{\lambda y} & \text{for } y < 0 \\ 1 - \frac{\lambda}{\lambda+\mu} e^{-\mu y} & \text{for } y \geq 0 . \end{cases} \tag{4.5}$$

Thus $g(y) = \frac{d}{dy} P\{X_n \leq y\}$ exists for all y and we write $P\{X_n \in dy\} = g(y)dy$ where

$$g(y) = \begin{cases} g_-(y) = \frac{\lambda\mu}{\lambda+\mu} e^{\lambda y} , & y < 0 \\ g_+(y) = \frac{\lambda\mu}{\lambda+\mu} e^{-\mu y} , & y \geq 0 . \end{cases} \tag{4.6}$$

Now to evaluate $f_1(x)$ we have

$$f_1(x) = \int_{[0,\infty)} P(x,\ dy)\ f(y) = \int_{[0,\infty)} yP\{(W_n + X_{n+1})^+ \in dy | W_n = x\}$$

$$= \int_{(0,\infty)} yP\{x + X_{n+1} \in dy\} = \int_{(0,\infty)} yg(y-x)dy$$

$$= \int_{y=0}^{x} yg_-(y-x)dy + \int_{y=x}^{\infty} yg_+(y-x)dy \quad .$$

Evaluation of these integrals is straightforward and we find that

$$f_1(x) = x + \frac{\lambda-\mu}{\mu\lambda} + \frac{\mu}{\lambda(\lambda+\mu)}\ e^{-\lambda x} \quad . \tag{4.7}$$

To evaluate $f_2(x)$ we must compute the integral

$$f_2(x) = \int_{[0,\infty)} P(x,dy)\ f_1(y) \tag{4.8}$$

$$= P(x,\ \{0\})\ f_1(0) + \int_{(0,\infty)} f_1(y)\ g(y-x)dy$$

$$= P\{X_{n+1} \leq -x\}\ f_1(0) + \int_{(0,\infty)} f_1(y)\ g(y-x)dy \quad .$$

It is finally found that

$$f_2(x) = f_1(x) + \frac{\lambda-\mu}{\mu\lambda} + \left(\frac{\mu}{\lambda+\mu}\right)^2 e^{-\lambda x}\left(\frac{1}{\lambda} + \frac{1}{\lambda+\mu} + x\right) \quad . \tag{4.9}$$

It is possible in this case to calculate exactly the covariance matrix Σ_2 so that we can obtain exact results for the variance reductions.

The calculations are very long and tedious and so will be omitted. The basic approach is to use stationary process theory; i.e., we assume that $\{W_n,\ n \geq 0\}$ is a stationary process. We can then obtain the central limit theorem given in Proposition (3.35). The covariance matrix V for this central limit theorem is given in (3.36). Since the central limit theorem given in Proposition (2.21) (with its covariance matrix $\Sigma/E(\tau_1)$) is still valid no matter what the initial distribution of the $\{W_n\}$ process is, it must be the case that the two seemingly different matrices are equal. Therefore, $v_{ij} = \sigma_{ij}/E(\tau_1)$. We can now evaluate the terms v_{ij} by performing calculations on

$$(4.10) \qquad R(z_1, z_2, s) = \sum_{n=0}^{\infty} E\left[e^{-z_1 W_0 - z_2 W_n}\right] s^n \ ,$$

the generating function of the joint Laplace transform of the (stationary) waiting time process. $R(z_1, z_2, s)$ has been calculated by Blomqvist (1967). Table 7 gives the calculated variance reductions in the M/M/1 queue for different values of ρ. Again we obtain substantial variance reductions although the method is somewhat less effective for high values of ρ.

To determine how well the method performs in actual simulation two values of ρ, $\rho = .5$ and $\rho = .9$, were selected for simulation. The simulation run lengths were 200,000 and 90,000 cycles for $\rho = .5$ and .9 respectively. The expected number of customers simulated for these two runs are 400,000 and 900,000. The random number generator described in Learmonth and Lewis (1973) was used. Statistical properties of this generator are reported in Learmonth and Lewis (1974).

Each run was broken into R independent replications of C cycles/replication for several different values of R and C (RC = total number of cycles simulated). For example in the $\rho = .5$ run we had $R = 200, 100, 50,$ and 1 with their corresponding values of $C = 1000, 2000, 4000,$ and $200{,}000$. For each replication we formed point estimates for the various parameters of interest. The figures reported in Tables 8, 9, 11, and 12 are then the sample averages of the point estimates taken over the R independent replications. Approximate 95% confidence intervals for each parameter (formed in the usual manner using the R i.i.d. replications) are given directly below the point estimator. For example, in Table 8, for $R = 200$ and $C = 1000$ a 95% confidence interval for $r = E(W) = 1000$ based on the estimator $\hat{r}_0$ is $(1.015 - .016, 1.015 + .016)$. In these tables we include the dependence of β^* on k by letting β_k^* denote the vector of optimal multipliers when combining the estimators $\hat{r}_0, \ldots, \hat{r}_k$, and we let

$$\hat{r}_{\hat{\beta}_k^*} = \sum_{\nu=0}^{k} \hat{\beta}_k^*(\nu)\, \hat{r}_\nu \tag{4.11}$$

where $\hat{\beta}_k^*(\nu)$ is our point estimate for $\beta_k^*(\nu)$.

On each replication 90, 95 and 99% confidence intervals were formed using each of the point estimators $\hat{r}_0$, $\hat{r}_1$, $\hat{r}_2$, $\hat{r}_{\hat{\beta}_1^*}$ and $\hat{r}_{\hat{\beta}_2^*}$ and their appropriate estimated variance terms. The fraction of those confidence intervals that actually contain r are given in Tables 10 and 13. If valid confidence intervals are being formed, these fractions (called

coverages) should coincide (approximately) with the type of confidence interval being formed; i.e., we expect

$$\frac{\text{number of } 100(1-\alpha)\% \text{ confidence intervals containing } r}{\text{number of } 100(1-\alpha)\% \text{ confidence intervals formed}} \approx 1-\alpha \ .$$

In these runs only "classical" point estimates for r and the covariance matrix were formed. The methods described in this paper are still valid using other types of point and interval estimates such as jackknife estimates. The reader should consult Iglehart (1975) for more details on these types of estimates.

The point estimates for $\underset{\sim}{\beta}_k^*$ were formed using (3.15). Notice that to estimate $\underset{\sim}{\beta}_k^*$ we use the point estimates for Σ_k based on the entire run length of the simulation. It has been suggested in Lavenberg, Moeller and Sauer (1977) that if the optimal coefficients in control variable schemes are estimated from a relatively short run and then a long run is made independently, with the coefficients set at these estimated levels, that an increase in coverage can be expected. However the cost of doing so is to decrease the amount of variance reduction obtained, so that this strategy was not adopted. From Tables 10 and 13 it can be seen that as C, the number of cycles/replication, increases (and thus R decreases) the coverage probabilities generally increase. This behavior has also been observed by Law (1976) in the context of studying the trade-offs in choosing the number of replications and run length per replication of a simulation. It should be noted that in every case for our longest runs ($R = 1$) coverage of $E(W)$ was achieved using the estimated coefficients.

The relatively low coverages for shorter runs (small C) can be explained by observing in Tables 9 and 12 that $\sigma_k^2(\beta_k^*)$ is underestimated. This causes the confidence intervals to be too short thus resulting in lowered coverage. The main reason that $\sigma_k^2(\beta_k^*)$ is underestimated is due to inaccurate estimation of Σ_k (see Tables 9 and 12).

Because of this behavior the simulator must take care that the run length is not too short. By having too short a run length the simulator may be placing unjustified confidence in his/her estimates. If the run length is adequate however the method is capable of producing tight confidence intervals with good coverage properties. The techniques developed by Lavenberg and Sauer (1977) for determining run lengths are applicable and may be of practical value in this area.

In order to use this method we must evaluate the functions $f_0(W_n)$, $f_1(W_n)$ and $f_2(W_n)$ for each customer n. To get a measure of the computational savings (if any) of the method it is important to determine how much extra work is being done for each customer. From (4.10) and (4.12) it is seen that to evaluate f_1 and f_2 one exponential and several multiplications and additions must be computed for each customer. For $\rho = .5$ we estimated (from CPU times of shorter runs) that on the average each customer using multiple estimates requires 5/3 as much CPU time as a run using no variance reduction technique. Since we need only simulate .0327 times as many customers to get confidence intervals of equal length (see Table 7) our computational savings, which is defined as the ratio of CPU times needed to obtain equally accurate estimates, is .0545 (= 5/3 × .0327). For $\rho = .9$ the work per customer is increased by less than a factor of 5/3. This is because for a fixed number of

customers we must evaluate f_2 and f_3 the same number of times for the different values of ρ, but we experience fewer cycles with the higher value of ρ. We therefore spend less time updating estimates of Σ. Thus for $\rho = .9$ our computational savings is at least $.547$ $(= 5/3 \times .328)$. On the other hand if we fix the CPU time to be the same for each method what is the statistical savings (defined to be the ratio of confidence interval lengths for equal run times)? Suppose for a specified CPU time we can simulate N_1 customers using no variance reduction technique (method 1) and N_2 customers using multiple estimates (method 2). Since each method 2 customer requires $5/3$ as much CPU time as a method 1 customer we must have $N_2 = 3/5\ N_1$. The ratio of the lengths of the confidence intervals is then

$$\frac{\sigma_0/N_1^{1/2}}{\sigma_2(\beta^*)/N_2^{1/2}} = \frac{\sigma_0}{\sigma_2(\beta^*)} \times (5/3)^{1/2} \quad .$$

This ratio is $.23$ and $.74$ for $\rho = .5$ and $.9$ respectively.

TABLE 7

Calculated Variance Reductions for the Waiting Time Process in an M/M/1 Queue: $r = E(W)$

($\mu = 1$, $\lambda = \rho$, $E(W) = \rho/(\mu-\lambda)$)

ρ	$E(W)$	σ_0^2	R_1^2 R_1	R_2^2 R_2
.10	.1111	.375	.0070 .0838	.0001 .0008
.20	.250	1.39	.0265 .1628	.0009 .0306
.30	.429	3.96	.0568 .2383	.0045 .0672
.40	.667	10.6	.0968 .3111	.0137 .1173
.50	1.00	29.0	.1457 .3817	.0327 .1808
.60	1.50	88.5	.2029 .4505	.0664 .2577
.70	2.33	336	.2678 .5175	.1214 .3485
.80	4.00	1,976	.3397 .5828	.2055 .4533
.90	9.00	35,901	.4175 .6461	.3280 .5727
.95	19.00	607,601	.4582 .6769	.4072 .6381
.99	99.00	3.96×10^8	.4904 .7003	.4686 .6845

TABLE 8

Simulation Results for Waiting Time Process in M/M/1 Queue With $\rho = .5$, Point Estimates and 95% Confidence Intervals

Parameter	True Value	R = 200 C = 1000	R = 100 C = 2000	R = 50 C = 4000	R = 1 C = 200,000
$\hat{r}_0$	1.000	1.015 .016	1.016 .017	1.018 .015	1.018
$\hat{r}_1$	1.000	1.013 .013	1.014 .014	1.015 .012	1.015
$\hat{r}_2$	1.000	1.011 .011	1.011 .011	1.012 .010	1.013
$\hat{r}_{\hat{\beta}_1^*}$	1.000	.999 .006	1.000 .007	1.003 .006	1.003
$\hat{r}_{\hat{\beta}_2^*}$	1.000	.996 .003	.998 .003	.999 .003	1.001
R_1^2	.146	.106 .005	.116 .007	.125 .007	.135
R_1	.382	.321 .008	.337 .010	.352 .011	.367
R_2^2	.033	.015 .001	.018 .002	.022 .003	.026
R_2	.181	.115 .005	.129 .007	.144 .008	.160
$\beta_1^*(0)$	-3.645	-3.572 .094	-3.635 .106	-3.692 .107	-3.727
$\beta_1^*(1)$	4.645	4.572 .094	4.635 .106	4.692 .107	4.727
$\beta_2^*(0)$	5.396	4.587 .222	4.866 .266	5.123 .274	5.343
$\beta_2^*(1)$	-16.700	-14.648 .573	-15.365 .686	-16.024 .710	-16.585
$\beta_2^*(2)$	12.304	11.061 .351	11.498 .421	11.901 .436	12.242

TABLE 9

Point Estimates and 95% Confidence Intervals for Variances and Covariances, M/M/1 Simulation, $\rho = .5$

Parameter	True Value	R = 200 C = 1000	R = 100 C = 2000	R = 50 C = 4000	R = 1 C = 200,000
σ_0^2	29.0	30.23 2.53	30.56 2.67	30.83 2.34	30.98
σ_{01}	23.67	24.67 2.21	24.95 2.34	25.18 2.05	25.31
σ_{02}	19.48	20.30 1.93	20.53 2.04	20.73 1.80	20.83
σ_1^2	19.48	20.30 1.94	20.53 2.05	20.73 1.81	20.84
σ_{12}	16.14	16.79 1.69	16.99 1.79	17.16 1.59	17.25
σ_2^2	13.44	13.95 1.48	14.12 1.56	14.27 1.39	14.34
$\sigma_1^2(\underset{\sim}{\beta}_1^*)$	4.23	3.46 .40	3.74 .46	3.98 .46	4.18
$\sigma_2^2(\underset{\sim}{\beta}_2^*)$	.948	.505 .079	.603 .104	.699 .119	.792

TABLE 10

Probability of Coverage for 90, 95 and 99% Confidence Intervals for E(W) in M/M/1 Queue, $\rho = .5$

Estimator	Type of C.I.	R = 200 C = 1000	R = 100 C = 2000	R = 50 C = 4000	R = 1 C = 200,000
$\hat{r}_0$	.90	.89	.92	.92	0.00
	.95	.95	.96	.98	0.00
	.99	.99	.98	1.00	1.00
$\hat{r}_1$	.90	.89	.90	.92	0.00
	.95	.95	.94	.96	0.00
	.99	.98	.98	1.00	1.00
$\hat{r}_2$	.90	.89	.89	.94	0.00
	.95	.94	.94	.96	0.00
	.99	.98	.98	.98	1.00
$\hat{r}_{\hat{\beta}_1^*}$	.90	.81	.84	.92	1.00
	.95	.87	.88	.94	1.00
	.99	.93	.93	.96	1.00
$\hat{r}_{\hat{\beta}_2^*}$	.90	.69	.76	.78	1.00
	.95	.74	.79	.90	1.00
	.99	.84	.86	.98	1.00

TABLE 11

Simulation Results for Waiting Time Process in M/M/1 Queue With $\rho = .9$, Point Estimates and 95% Confidence Intervals

Parameter	True Value	R = 60 C = 1500	R = 30 C = 3000	R = 15 C = 6000	R = 1 C = 90,000
$\hat{r}_0$	9.000	8.841 .283	8.863 .274	8.891 .210	8.900
$\hat{r}_1$	9.000	8.842 .281	8.864 .271	8.892 .209	8.901
$\hat{r}_2$	9.000	8.843 .279	8.865 .270	8.892 .207	8.901
$\hat{r}_{\hat{\beta}_1^*}$	9.000	8.875 .217	8.929 .225	8.988 .278	8.950
$\hat{r}_{\hat{\beta}_2^*}$	9.000	8.827 .187	8.899 .210	8.946 .253	8.908
R_1^2	.417	.312 .017	.339 .023	.355 .021	.384
R_1	.646	.555 .016	.579 .020	.595 .018	.620
R_2^2	.328	.218 .017	.245 .022	.263 .021	.296
R_2	.573	.461 .018	.491 .023	.511 .020	.544
$\beta_1^*(0)$	-105.1	-89.9 7.2	-93.5 8.3	-95.7 8.9	-96.5
$\beta_1^*(1)$	106.1	90.9 7.2	94.5 8.3	96.7 8.9	97.5
$\beta_2^*(0)$	462	370 38	392 43	401 43	408
$\beta_2^*(1)$	-1050	-846 84	-895 96	-916 95	-930
$\beta_2^*(2)$	589	477 46	504 52	516 53	523

TABLE 12

Point Estimates and 95% Confidence Intervals for Variances and Covariances, M/M/1 Simulation, $\rho = .9$

Parameter	True Value	R = 60 C = 1500	R = 30 C = 3000	R = 15 C = 6000	R = 1 C = 90,000
σ_0^2	35,901	26,649 7,498	27,809 8,140	28,636 8,147	28,601
σ_{01}	35,704	26,474 7,473	27,631 8,115	28,454 8,123	28,420
σ_{02}	35,509	26,302 7,449	27,454 8,090	28,275 8,099	28,242
σ_1^2	35,509	26,302 7,449	27,455 8,091	28,276 8,099	28,242
σ_{12}	35,315	26,131 7,425	27,281 8,066	28,098 8,076	28,065
σ_2^2	35,123	25,962 7,401	27,107 8,042	27,923 8,053	27,890
$\sigma_1^2(\beta_1^*)$	14,988	8,142 1,989	9,454 2,817	10,348 3,553	10,977
$\sigma_2^2(\beta_1^*)$	11,777	5,743 1,380	6,917 2,112	7,801 2,905	8,452

TABLE 13

Probability of Coverage for 90, 95 and 99% Confidence Intervals for E(W) in M/M/1 Queue, $\rho = .9$

Estimator	Type of C.I.	R = 60 C = 1500	R = 30 C = 3000	R = 15 C = 6000	R = 1 C = 90,000
$\hat{r}_0$	.90	.88	.93	1.00	1.00
	.95	.92	.93	1.00	1.00
	.99	.93	.96	1.00	1.00
$\hat{r}_1$	.90	.88	.93	1.00	1.00
	.95	.92	.93	1.00	1.00
	.99	.93	.96	1.00	1.00
$\hat{r}_2$	.90	.88	.93	1.00	1.00
	.95	.92	.93	1.00	1.00
	.99	.93	.96	1.00	1.00
$\hat{r}_{\hat{\beta}_1^*}$	.90	.75	.83	.80	1.00
	.95	.85	.90	.87	1.00
	.99	.90	1.00	1.00	1.00
$\hat{r}_{\hat{\beta}_2^*}$	.90	.72	.77	.80	1.00
	.95	.80	.87	.93	1.00
	.99	.89	.97	1.00	1.00

(4.12) EXAMPLE. Waiting Time Process in an M/M/2 Queue.

We now turn to a stochastic process which has a much more complicated structure than any of the previous processes. For this example we are interested in estimating the expected stationary waiting time in a GI/G/c queue. The number of servers in the queue, c, is assumed to be greater than 1. To apply the method we need to find an underlying Markov process.

Following Kiefer and Wolfowitz (1955) and (1956) we let W_{ni} be the workload of the server with the ith lightest workload just prior to the arrival of the nth customer and let $\underset{\sim}{W}_n = (W_{n1}, \ldots, W_{nc})$. Again let S_n denote the service time of the nth customer and A_{n+1} denote the time between the arrival of the nth and (n+1)st customer. Let $\underset{\sim}{X}_{n+1}$ be a vector with c components defined by

$$\underset{\sim}{X}_{n+1} = (S_n - A_{n+1}, -A_{n+1}, \ldots, -A_{n+1}) . \tag{4.13}$$

A recursion for $\underset{\sim}{W}_n$ can be defined by (assuming the system is initially empty)

$$\underset{\sim}{W}_0 = (0, \ldots, 0) \tag{4.41}$$

$$\underset{\sim}{W}_{n+1} = F[(\underset{\sim}{W}_n + \underset{\sim}{X}_{n+1})^+] , \qquad n \geq 0$$

where F is the function from $\mathbb{R}^c$ to $\mathbb{R}^c$ which arranges components in increasing order. Notice that (4.4) is a spcial case of (4.41) when c = 1. The waiting time of the nth customer is then W_{n1}. If we assume that

$\{S_n,\ n \geq 0\}$ are i.i.d. with mean μ^{-1} and distribution function G and that $\{A_n,\ n \geq 1\}$ are i.i.d. with mean λ^{-1} and distribution function H, it is easy to see that $\{\underset{\sim}{W}_n,\ n \geq 0\}$ is a Markov process with state space $E = \{\underset{\sim}{x} \in \mathbb{R}^c : 0 \leq x_1 \leq x_2 \leq \cdots \leq x_c\}$. Let $\rho = \lambda/(\mu c)$. It is known that if $\rho < 1$ and if $P\{A_{n+1} > S_n\} > 0$ then there exists an infinite number of indices n such that $\underset{\sim}{W}_n = (0, \ldots, 0)$ and that the expected time between such indices is finite (see Whitt (1972)). We therefore pick $\underset{\sim}{x}_0 = (0, \ldots, 0)$ as our return state. From the above we know that there exists a $\underset{\sim}{W} = (W_1, \ldots, W_c)$ such that $\underset{\sim}{W}_n \Rightarrow \underset{\sim}{W}$. We shall be interested in estimating $r = E(W_1)$, the expected stationary waiting time. Let $f: E \to \mathbb{R}^+$ be defined by $f(x_1, \ldots, x_c) = x_1$, then $r = E(f(\underset{\sim}{W}))$.

In order to apply the method it is necessary to calculate $P^\nu f$ for $\nu = 1, \ldots, k$. Unfortunately even in such a simple case as the M/M/2 queue calculation of the transition function P is quite tedious (although it is possible to evaluate). For this reason we choose $k = 1$ since evaluation of Pf is relatively straightforward;

$$(4.15) \qquad Pf(\underset{\sim}{w}) = E(W_{n+1,1} \mid \underset{\sim}{W}_n = (w_1, w_2, \ldots, w_c))$$

$$= \int_{C_1} (w_1+s-t)\, dH(t)\, dG(s) + \sum_{h=2}^{c} \int_{C_h} (w_k-t)\, dH(t)\, dG(s)$$

where $C_1 = \{(s,t):\ 0 \leq w_1+s-t \leq w_j-t \text{ for } j \geq 2\}$ and $C_h = \{(s,t): 0 \leq w_h-t \leq w_j-t \text{ for } j \geq 2 \text{ and } w_h-t < w_1+s-t\}$ for $h \geq 2$. Notice though that C_h is empty for $h > 2$ since we must have $w_2 \leq w_h$

in which case $w_2 - t \leq w_h - t$ (in case $w_2 = w_h$ assign the ambiguous points to C_2 rather than C_h). Thus for all values of $c \geq 2$ we have

$$(4.16) \qquad Pf(\underset{\sim}{w}) = \int_{C_1} (w_1 + s - t)\, dH(t)\, dG(s) + \int_{C_2} (w_2 - t)\, dH(t)\, dG(s) .$$

Again the evaluation of these integrals is straightforward for the M/M/c queue. In this case we find that

$$(4.17) \qquad f_1(\underset{\sim}{w}) = w_1 + \left(\frac{1}{\lambda} - \frac{1}{\mu}\right) + \frac{\mu}{\lambda(\lambda+\mu)} e^{-\lambda w_1} + e^{\mu(w_1 - w_2)} \left(\frac{e^{-\lambda w_2}}{\lambda+\mu} - \frac{1}{\mu}\right) .$$

For this process the covariance matrix Σ is unknown so that actual variance reductions cannot be calculated. We again simulated the process for $\rho = .5$ and $\rho = .9$. Our run lengths were 100,000 cycles (300,000 customers) for $\rho = .5$ and 40,000 cycles (760,000 customers) for $\rho = .9$. The results of these simulations are reported in Tables 14 to 17. As in the M/M/1 case the coverage probabilities generally increase with the run length. For a fixed value of ρ the variance reduction obtained is less for the M/M/2 queue than for the M/M/1 queue. This is probably due to the fact that the underlying stochastic process for M/M/2 is quite a bit more complicated than for M/M/1.

Notice that to evaluate the function f_1 we must compute three exponentials. Since this must be done for each customer we are substantially increasing the CPU time required for the simulation. In fact for $\rho = .9$ we estimate that each customer requires 2.25 as much CPU time as straightforward simulation. Since R_1^2 (see Table 16) is greater than .5

it must be concluded that in this case the method is not computationally efficient. By this we mean that for a fixed amount of CPU time we can get more accurate estimates by not using the variance reduction technique. For this reason the method is not recommended for the heavily loaded GI/G/c queue ($c > 1$) unless the value of k can be increased and the functions $f_1, \ldots, f_k$ can be evaluated cheaply.

TABLE 14

Simulation Results for Waiting Time Process in M/M/2 Queue With $\rho = .5$, Point Estimates and 95% Confidence Intervals

Parameter	True Value	R = 100 C = 1000	R = 50 C = 2000	R = 25 C = 4000	R = 1 C = 100,000
$\hat{r}_0$	.6667	.6651 .0212	.6672 .0181	.6677 .0182	.6683
$\hat{r}_1$	.6667	.6664 .0184	.6682 .0159	.6686 .0159	.6692
$\hat{r}_{\hat{\beta}_1^*}$	.6667	.6607 .0101	.6683 .0100	.6702 .0108	.6726
R_1^2		.3029 .0153	.2983 .0174	.3052 .0197	.3207
R_1		.5457 .0141	.5434 .0156	.5508 .0171	.5663
σ_0^2		26.39 4.83	27.00 4.85	27.14 4.50	27.35
σ_{01}		22.90 4.46	23.44 4.48	23.56 4.15	23.75
σ_2^2		20.10 4.11	20.57 4.15	20.67 3.84	20.85
$\sigma_1^2(\beta^*)$		7.38 1.35	7.94 1.58	8.36 1.71	8.77

TABLE 15

Probability of Coverage for 90, 95 and 99% Confidence Intervals for E(W) in M/M/2 Queue, $\rho = .5$

Estimator	Type of C.I.	R = 100 C = 1000	R = 50 C = 2000	R = 25 C = 4000	R = 1 C = 100,000
$\hat{r}_0$	.90	.80	.90	.88	1.00
	.95	.85	.92	.96	1.00
	.99	.92	.96	.96	1.00
$\hat{r}_1$	.90	.80	.90	.88	1.00
	.95	.84	.92	.96	1.00
	.99	.93	.96	.96	1.00
$\hat{r}_{\hat{\beta}_1^*}$	.90	.84	.92	.80	1.00
	.95	.87	.94	.92	1.00
	.99	.96	.96	1.00	1.00

TABLE 16

Simulation Results for Waiting Time Process in M/M/2 Queue with $\rho = .9$, Point Estimates and 95% Confidence Intervals

Parameter	True Value	R = 32 C = 1250	R = 16 C = 2500	R = 8 C = 5000	R = 1 C = 400,000
$\hat{r}_0$	8.526	8.684 .481	8.713 .500	8.728 .408	8.745
$\hat{r}_1$	8.526	8.686 .479	8.714 .498	8.729 .406	8.746
$\hat{r}_{\hat{\beta}^*_1}$	8.526	8.772 .511	8.795 .417	8.904 .356	8.858
R_1^2		.536 .044	.559 .049	.556 .049	.577
R_1		.727 .030	.745 .033	.744 .033	.760
σ_0^2		52,975 24,283	55,206 25,518	57,671 23,670	57,720
σ_{01}		52,788 24,232	55,015 25,465	57,473 23,627	57,523
σ_2^2		52,604 24,182	54,827 25,412	57,278 23,584	57,328
$\sigma_1^2(\beta^*)$		24,055 9,063	28,660 12,112	31,928 13,555	33,326

TABLE 17

Probability of Coverage for 90, 95 and 99% Confidence Intervals for E(W) in M/M/2 Queue, $\rho = .9$

Estimator	Type of C.I.	R = 32 C = 1250	R = 16 C = 2500	R = 8 C = 5000	R = 1 C = 40,000
$\hat{r}_0$	.90	.84	.94	1.00	1.00
	.95	.88	.94	1.00	1.00
	.99	.91	.94	1.00	1.00
$\hat{r}_1$	.90	.84	.94	1.00	1.00
	.95	.88	.94	1.00	1.00
	.99	.91	.94	1.00	1.00
$\hat{r}_{\hat{\beta}_1^*}$	.90	.84	.88	.88	1.00
	.95	.84	.94	1.00	1.00
	.99	.91	.94	1.00	1.00

5. Conclusions

In this paper a new variance reduction technique for a wide class of stochastic processes is proposed and tested. The method differs from most other control variable methods in that the means of the control variables do not need to be known explicitly. The method is capable of producing substantial variance reductions. Because the method requires additional computations to be done both before and during the simulation, care must be taken so that the method is used only when it is computationally advantageous to do so, that is it should only be used when for a fixed amount of computer time more accurate estimates can be obtained by using the method than by not using it. In the case of Markov chains it is likely that the method will be most effective when the transition matrix of the process is sparse, in which case the preliminary calculations can be carried out with relative ease. It is for this type of process that the method is recommended.

Acknowledgment

The author wishes to thank Professor Donald L. Iglehart for his valuable suggestions during this research. The author would also like to thank Gail Lemmond for her expert typing of the manuscript. This research was supported by National Science Foundation grant number MCS75-23607 and Office of Naval Research contract N00014-76-C-0578.

BIBLIOGRAPHY

[1] BILLINGSLEY, P. (1968). Convergence of Probability Measures. John Wiley, New York.

[2] BLOMQVIST, N. (1967). The covariance function of the M/G/1 queuing system. Scand. Actuar. J. 50, 157-174.

[3] CHUNG, K.L. (1967). Markov Chains with Stationary Transition Probabilities, 2nd edn. Springer-Verlag, Berlin.

[4] ÇINLAR, E. (1975). Introduction to Stochastic Processes. Prentice-Hall, Inc., Englewood Cliffs, N.J.

[5] CRANE, M.A. and IGLEHART, D.L. (1974a). Simulating stable stochastic systems, I: General multi-server queues. J. Assoc. Comput. Mach. 21, 103-113.

[6] CRANE, M.A. and IGLEHART, D.L. (1974b). Simulating stable stochastic systems, II: Markov chains. J. Assoc. Comput. Mach. 21, 114-123.

[7] CRANE, M.A. and IGLEHART, D.L. (1975). Simulating stable stochastic systems, III: Regenerative processes and discrete-event simulations. Operat. Res. 23, 33-45.

[8] DOOB, J.L. (1953). Stochastic Processes. John Wiley, New York.

[9] GAVER, D.P. and SHEDLER, G.S. (1971). Control variable methods in the simulation of a model of a multiprogrammed computer system. Naval Res. Logist. Quart. 18, 435-450.

[10] GAVER, D.P. and THOMPSON, G.L. (1973). Programming and Probability Models in Operations Research. Brooks/Cole Pub. Co., Monterey, Calif.

[11] HAMMERSLEY, J.M. and HANDSCOMB, D.C. (1964). Monte Carlo Methods. Methuen and Co. Ltd., London.

[12] HORDIJK, A., IGLEHART, D.L. and SCHASSBERGER, R. (1976). Discrete time methods for simulating continuous time Markov chains. Adv. Appl. Prob. 8, 772-788.

[13] IGLEHART, D.L. (1975). Simulating stable stochastic systems, V: Comparison of ratio estimators. Naval Res. Logist. Quart. 22, 553-565.

[14] IGLEHART, D.L. (1971). Functional limit theorems for the queue GI/G/1 in light traffic. Adv. Appl. Prob. 3, 269-281.

[15] IGLEHART, D.L. and LEWIS, P.A.W. (1976). Variance reduction for regenerative simulations, I: Internal control and stratified sampling for queues. Technical Report No. 86-22, Control Analysis Corp., Palo Alto, Calif.

[16] KARLIN, S. and TAYLOR, H.M. (1975). A First Course in Stochastic Processes, 2nd edn. Academic Press, New York.

[17] KIEFER, J. and WOLFOWITZ, J. (1955). On the theory of queues with many servers. Trans. Amer. Math. Soc. 78, 1-18.

[18] KIERFER, J. and WOLFOWITZ, J. (1956). On the characteristics of the general queueing process with applications to random walks. Ann. Math. Statist. 27, 147-161.

[19] LAVENBERG, S.S. (1974). Efficient estimation of work-rates in closed queueing networks. Proceedings in Computational Statistics. Physica Verlag, Vienna, 353-362.

[20] LAVENBERG, S.S., MOELLER, T.L. and SAUER, C.H. (1977). Concomitant control variables applied to the regenerative simulation of queueing systems. IBM Research Report RC 6413, Yorktown Heights, New York.

[21] LAVENBERG, S.S. MOELLER, T.L. and WELCH, P.D. (1977). Control variables applied to the simulation of queueing models of computer systems. Computer Performance. North Holland Publishing Co., Amsterdam, 459-467.

[22] LAVENBERG, S.S. and SAUER, C.H. (1977). Sequential stopping rules for the regenerative method of simulation. IBM Research Report RC 6412, Yorktown Heights, New York.

[23] LAVENBERG, S.S. and SHEDLER, G.S. (1975). Derivation of confidence intervals for work rate estimators in a closed queueing network. SIAM J. Comput. 4, 108-124.

[24] LAW, A.M. (1976). Confidence intervals in discrete event simulation: A comparison of replication and batch means. Technical Report 76-13, Department of Industrial Engineering, Madison, Wisconsin.

[25] LEARMONTH, G.P. and LEWIS, P.A.W. (1973). Naval Postgraduate School random number generator package LLRANDOM. Naval Postgraduate School Report NP555Lw73061A, Monterey, Calif.

[26] LEARMONTH, G.P. and LEWIS, P.A.W. (1974). Statistical tests of some widely used and recently proposed uniform random number generators. Proc. Seventh Conference on Computer Science and Statistics, Western Periodicals Co., North Hollywood, Calif.

[27] OREY, S. (1971). Limit Theorems for Markov Chain Transition Probabilities. Van Nostrand Reinhold Co., London.

[28] WHITT, W. (1972). Embedded renewal processes in the GI/G/s queue. J. Appl. Prob. 9, 185-191.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 42	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Variance Reduction Techniques for the Simulation of Markov Processes, I: Multiple Estimates		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) PHILIP HEIDELBERGER		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0578
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Operations Research Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (NR 042-343)
11. CONTROLLING OFFICE NAME AND ADDRESS Statistics and Probability Program Office of Naval Research (Code 436) Arlington, Virginia 20360		12. REPORT DATE October 1977 13. NUMBER OF PAGES 73
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release: distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

SIMULATION, VARIANCE REDUCTION, MARKOV PROCESS

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

see attached

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Variance Reduction Techniques for the Simulation of Markov Processes, I: Multiple Estimates, by Philip Heidelberger

Let $\{X_n, n \geq 0\}$ be a regenerative Markov process with state space E, n-step transition matrix P^n, and stationary distribution π. Let $r = \pi f$. A detailed analysis of a method for reducing the variance of simulation estimates for r is presented. Let $f_\nu = P^\nu f$ and $\hat{x}_\nu(N) = \sum_{n=0}^{N} f_\nu(X_n)/(N+1)$ for $\nu = 0, \ldots, k$. Since $r = \pi f_\nu$, $\hat{x}_\nu(N) \to r$ a.s. as $N \to \infty$. Let $\hat{x}_\beta(N) = \sum_{\nu=0}^{k} \beta(\nu)\, \hat{x}_\nu(N)$. If $\sum_{\nu=0}^{k} \beta(\nu) = 1$ then $\hat{x}_\beta(N) \to r$ a.s. and we pick $\underset{\sim}{\beta} = \underset{\sim}{\beta}^*$ to minimize the asymptotic variance of $\hat{x}_\beta(N)$. The method can be extended to continuous time Markov chains, semi-Markov processes as well as certain types of stationary stochastic processes. Numerical results for a variety of simple queueing models are presented.